SOME NEW IDEAS AND TECHNIQUES FOR LEARNING SYSTEMS

By

CHARLES RICHARD REEVES

Bachelor of Science
University of Missouri at Rolla
Rolla, Missouri
1962

Master of Science
Oklahoma State University
Stillwater, Oklahoma
1966

Submitted to the Faculty of the Graduate College
of the Oklahoma State University
in partial fulfillment of the requirements
for the Degree of
DOCTOR OF PHILOSOPHY
May, 1970

SOME NEW IDEAS AND TECHNIQUES FOR

LEARNING SYSTEMS

Thesis Approved:

Bennett Basore
Thesis Adviser

Arthur M. Breipohl

Wm L Hughes

Kenneth A. McCollom

J Lloyd Felbo

D. Durham
Dean of the Graduate College

PREFACE

The basic purpose of this dissertation has been to shed some light on one of the less well researched and understood aspects of the phenomenon commonly known as "learning" and perhaps to stimulate interest in understanding the separate processes which are supposed to contribute to this phenomenon. It is hoped that with such an understanding the design and synthesis of more useful learning machines and adaptive systems will become possible.

The particular part of the learning process investigated here concerns the generation of new ideas and what might be called "insight" or "creativity". It has sometimes been thought that such capabilities may be possessed only by living systems, perhaps only by humans. Without bothering to argue the point, the contrary has been assumed in this investigation. This study is concerned not with <u>whether</u> such a process can be simulated on a machine but rather with <u>how</u> it can be done in general and in the particular simulated model proposed in the study.

The model used in the investigation is that of a communication system with a learning receiver. The receiver learns to recognize as distinct all of the different symbols sent by the transmitter in the presence of noise, without a priori knowledge of the number of symbols or the symbol features required for distinct recognition.

Attention is centered on the process whereby the receiver proposes new symbols for recognition which it supposes are being sent by the transmitter. The behavior of the learning receiver in this regard is

felt to demonstrate the feasibility of machine generation of relevant new ideas in a learning environment and gives some understanding of this basic part of the learning process.

I would like to express my sincere appreciation to Dr. Bennett L. Basore, my adviser, for his guidance and assistance throughout the study. Although he let it be a real "do-it-yourself" project, he was always ready with helpful suggestions and stimulating discussion. Thanks are also due to Dr. Arthur M. Breipohl for his interest, suggestions, and encouragement. The assistance of committee members Dr. William Hughes, Dr. Kenneth McCollom, and Dr. Leroy Folks is gratefully acknowledged.

A special thanks go to my wife, Sandee, for her encouragement, and help in the preparation of the final draft.

I would also like to thank Dixie Jennings for typing the final copy.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Learning and Inductive Inference

The simulation of human learning processes on computing machines generally takes the form of an attempt to perform inductive inference on a machine initially designed for deductive inference. The fact that programs have been written which play acceptably good games of chess and checkers, and which learn from their own experiences is evidence that some meaningful progress has been made in this direction (1,2).

Such game-playing programs generally use a process of testing various playing strategies against a backlog of past playing situations with known results in order to determine the relative merits of each strategy. This can be interpreted as showing that the programs do exhibit some capacity for inductive inference. The assignment of inductive probabilities, or credibilities, to the set of competing hypotheses, or strategies, using the store of empirical data and, perhaps, some a priori credibility for these hypotheses can certainly be said to be at least a part of the inductive process (3).

Up to this time there has been little evidence that machines using this principle are capable of "creative" work, and it is presently unknown whether truly creative machines will ever be feasible. Minsky (4) points out that limitations in this area may well not be limitations inherent in machines but rather in our present ability to construct and

program them. This idea seems to be prevalent at the present time but the controversy is by no means settled.

Present-day learning theory is far from being fully developed and there are many directions in which it may be extended. For example, the capability for generating completely new hypotheses, or ideas, is not provided in the present theory, although there is "room" in the theory for such a capability and many researchers feel that this is possible. At the present state of development most learning systems depend on the human system designer for a general formulation of the hypotheses. The learning systems can then "refine" those hypotheses to improve their performance.

The theory of inductive inference, as used in present-day learning machines, describes the behavior of the credibilities assigned to a set of competing hypotheses based on a growing body of empirical evidence. That is, it describes the manner in which a machine assigns credibilities to the set of hypotheses when the machine is presented with evidence which supports or denies the truth of the various hypotheses.

In order to relate the amount of learning to the credibilities associated with a set of hypotheses, let the amount of learning be considered "high" when unity credibility has been assigned to one hypothesis and zero credibility to all the others, or one hypothesis has been found to be the "law" governing the data. The amount of learning is "low" when all the hypotheses have equal credibility and there is great uncertainty as to which hypothesis is true. The usual entropy measure of the credibilities is often used to provide a qualitative measure of the amount of learning associated with the credibilities.

The R-function of Bakan (5) provided some initial insight into the

situations which affect the rate of learning. He showed that learning takes place rapidly when the empirical evidence is very closely associated with only one of the hypotheses, that is, the observed data could be the result of the truth of only one of the hypotheses. When the observed data may be the result of the truth of many of the hypotheses, learning takes place more slowly.

Basore (6), interpreting the R-function as a measure of the uniqueness of the data with respect to the hypotheses, investigated some of the theory of inductive inference from an information theoretical point of view using the R-function. In particular, he shows how the information flow through a learning system is affected by the state of learning in the system, and that this, in turn, can be expressed in terms of R.

Watanabe (3) gives a mathematically well-founded and comprehensive information theoretical analysis of inductive inference. His analysis, based on Bayes' theorem, shows that the expected value of the entropy of the credibilities must show a net decrease with increase in relevant empirical data, thus the system learns as more relevant empirical data are accumulated. He also gives a large body of methodological arguments justifying the model and discussing some of the more important points of inductive inference from an engineering point of view.

The Generation of Hypotheses and Learning

Fundamental to the theory at the present state of development is that only hypotheses which are actually formulated by the learning system may be judged. This seems natural enough until one comes to the situation in which none of the proposed hypotheses gives a good explanation of the observed data. What does one do when the conceptual model

applicable to an observed physical situation is not included in the set of hypotheses? Since the observed data must be "possible" in view of the proposed hypotheses, the data will tend to support those hypotheses from which the data could have arisen no matter how inappropriate those hypotheses are.

It would be very convenient if the theory included some method of detecting such situations and proposing new hypotheses which would better fit the data. Certainly it is not possible to derive hypotheses from data in a mathematical fashion since this is equivalent to reducing the whole problem of statistical inference to mathematical certainty. Then it must be left up to the learning system itself to propose new hypotheses by whatever means it has at its disposal.

One of the most remarkable things about living learning systems is that they have a capability for organizing information in a fashion almost never matched in machine simulations. It is contended in this paper that the reason for this disparity is not so much the problem that the machines cannot be made large enough and fast enough to do the required computations, but rather that the living learning systems seem to possess capabilities which have not been identified in enough detail to permit simulation by a machine.

Gestalt psychology contends that there are various factors which organize, for instance, the visual field so that data received by the eyes is immediately perceived as having certain properties, without any need to consider all the possible properties that the data might possess. For example, elements in the visual field that are closest to each other tend to be perceived as groups; when more than one kind of element is present, those which are similar tend to form groups; and

lines which enclose a surface tend to be seen as a unit, or a "shape". Furthermore, it is contended that these perceived properties are somehow "natural", although previous experience helps in the formation of these concepts. In a general way, Gestalt theory promotes the belief that learning systems, or at least human beings, possess a faculty known as "insight" whereby new ideas, or hypotheses, are generated in a remarkably efficient manner.

This paper does not propose the construction of a machine having insight, or that the construction of such a machine is even possible. However, it is proposed to show how a learning system can be constructed that learns through its own experience in a fashion considerably more efficient than by completely random generation and testing of hypotheses, and that such a machine can exhibit some seemingly Gestalt actions.

There is at the present time no formal theory explaining the hypothesis generating phenomenon observed in living learning systems. The approach in this paper is exploratory in nature and, as will be explained later, there may be little value in attempting the formulation of a rigorous theory with the limited amount of data available to date.

A Typical Problem

As indicated, the identification of clusters of points in the visual field is a typically Gestalt phenomenon and, indeed, one that is difficult to simulate on a computer because the mechanism of identifying clusters is not well understood. For example, when a person is shown a set of points such as in Figure 1 and asked to identify any patterns in the data, one would probably reply that there are "obviously" three clusters of points like that indicated in Figure 2.

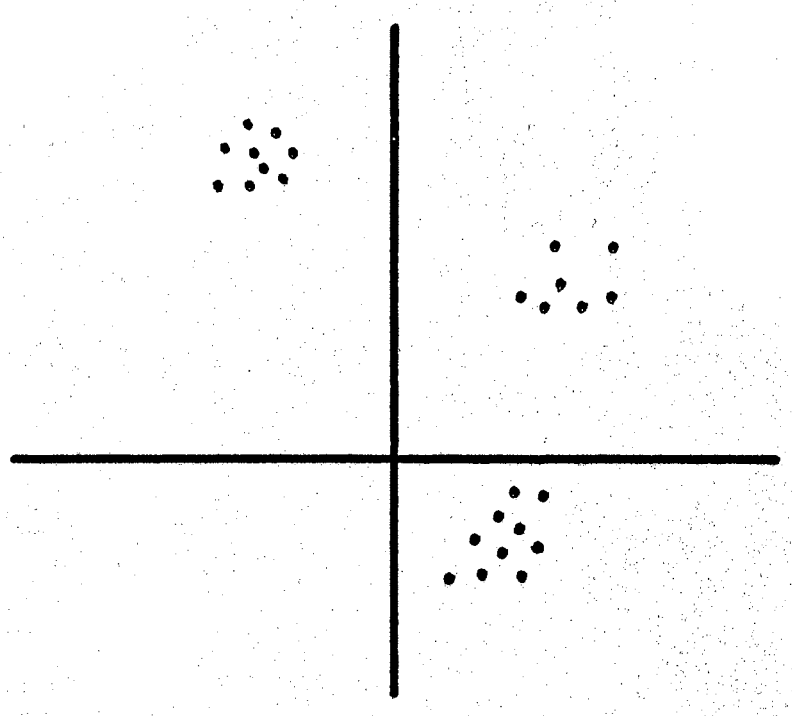

Figure 1. Example of Clustering Problem

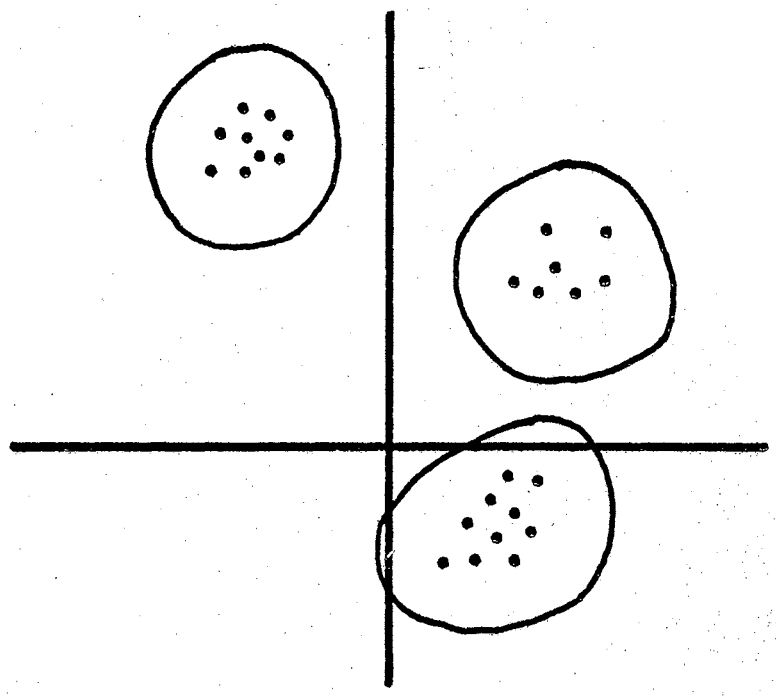

Figure 2. Possible Solution to First Clustering Problem

Actually, there are about 2.5×10^{12} different ways the points can be divided into three groups and, according to the theory of inductive inference at least, one would have to weigh all these different ways (hypotheses) to find the most likely one. It seems quite evident that this is not done. The solution is much more "natural".

Perhaps more obscure is the way in which the number of clusters, three in this case, is determined. One frequently proposed method of determining the best number of groups is to choose that number which minimizes some measure of the within-class variance of each group averaged over all the groups. Exactly how this measure is determined is not obvious since raising the number of groups always decreases the within-group variance, although the decrease becomes less and less with each increase in the number of groups. Perhaps one should choose the most "economical" number of groups in this respect.

There are, of course, many other acceptable ways of dividing the points, for instance, putting each point into its own group might seem like the best way to some. Generally, the answer might vary from person to person because of the difference in the way the problem is perceived by different individuals.

If the problem is changed only slightly the answers obtained may be strikingly different. If the points are arranged as in Figure 3, most answers would look something like Figure 4. It is immediately evident in this case that the similarity factor for each group is not closeness of points to each other but distance of the points from the origin.

In either case, a person's a priori knowledge, or belief, about just what is a pattern, how patterns can be expressed as groups of

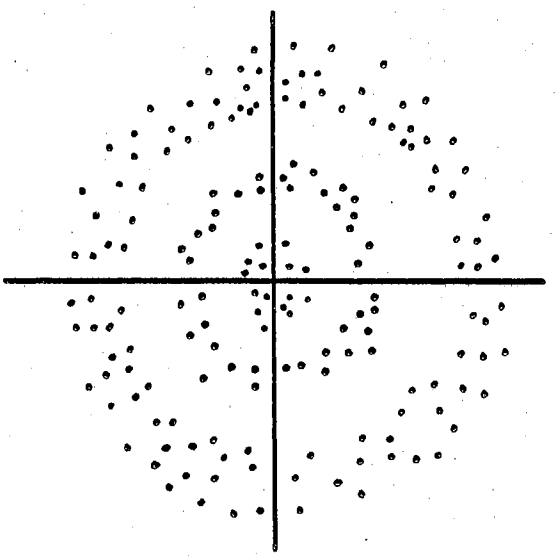

Figure 3. Example of Clustering Problem

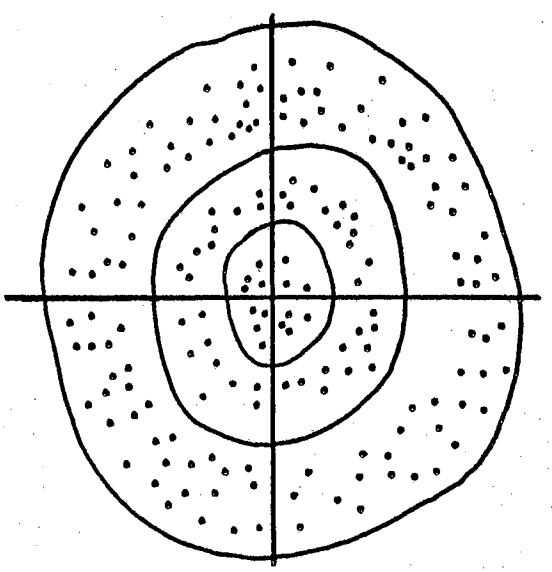

Figure 4. Possible Solution to Second Clustering Problem

points, and how many different groups there should be plays some role in obtaining the answer. In addition, the data itself may suggest remarkably different approaches to the problem. Finally, it seems certain that not all the possible hypotheses are tested, but that the process of solving the problem converges on the right answer, or at least an answer acceptable to the problem solver, much more rapidly, perhaps by a system of learning from a few trials. Probably some knowledge about the correct answer is gained from trials that yield unsatisfactory answers.

The Use of Data in Learning

Another interesting question to which very little attention has been devoted in the past is the question of how one knows what information is useful and/or necessary to solve a given problem. In the clustering problem there is only a small amount of information given to begin with, however, even some of that is probably discarded, or not noticed by the average person. For instance, the coordinates of the points with respect to specific axes seems to have no value in finding patterns among the points, and whatever information could be obtained from this source is probably disregarded.

In many situations there is more raw data available than can be economically handled and so one resorts to some form of preprocessing to strip off data believed to be redundant and irrelevant. Perhaps the simplest form of preprocessing is that of throwing away or disregarding part of the available data. In the clustering problem, the orientation of the points is disregarded because it seems to be irrelevant.

There is an important distinction to be made here between redundant

and irrelevant data. Redundancy as used here is a property of the data itself wherein the information is overspecified or repeated in a given body of data. On the other hand, irrelevancy is somehow a property of the use to which the data is put. Data may be either relevant or irrelevant depending on the particular problem it is used to solve. In the first clustering problem, the distance from each point to the origin is irrelevant; but in the second problem, that distance is the key to the particular solution given.

The point to be made by this discussion is that the way the data is handled depends greatly on the concept of the problem to be solved. In a situation where the problem is very general, such as the clustering problem, the solver's idea of what is expected in the solution affects the way the data is processed and, ultimately, the resulting answer.

In addition, it should be evident that the problem solver's concept of the problem may change as the solution is attempted, which leads to different handling of the data. The fact that the data itself may affect the way in which it is processed leads to the concept of a dynamic system with feedback which converges to a stable state when the system is "satisfied" with its own performance.

If a solution to the second clustering problem were attempted on the basis of having all the points in each group close together, the only possible answer would be to have all the points in one group - not a very satisfying result. Perhaps the fact that no good solution of this type is found causes the data to be reprocessed, whereupon the characteristic feature may be found to permit a satisfying answer to the problem.

A Model of a Learning System

In order to investigate some of these aspects of learning, it is proposed to study a model of a communication system with a learning receiver and to identify in its behavior some of the previously discussed phenomena. The system consists of a transmitter, a noise-inducing channel, and a self-evaluating, learning receiver synchronized with the transmitter.

The proposed model is one in which all the learning takes place at the receiver without benefit of a teacher to inform the receiver of its successes and failures. Thus, learning is based on the receiver's own evaluation of its performance according to some basic, fixed measure of goodness of performance. That is, the receiver attempts to adjust its parameters so that its own measure of satisfaction with its performance yields at least some minimum value. The criteria by which the receiver judges its performance is, unfortunately, fixed and this surely limits the ultimate degree of learning by the system. No claim is made that the system is creative, but merely that some aspects of learning may be observed in the system's behavior.

The receiver is given the task of recognizing which one of a finite set of symbols is sent by the transmitter during each time interval. The only information available to the reciver at the outset is statistical knowledge of the noise introduced into the received signal by the channel. In addition, the receiver is given the criteria by which it evaluates its own performance and some general "beliefs" about the type of environment in which it is to operate.

Since the receiver has no initial knowledge of the set of symbols sent by the transmitter, it must first form an estimate of the

transmitter symbol set and then "recognize" the received signals as particular representations of the estimated symbol set. The set of estimated symbols, called the receiver symbol set, is obtained by processing the received data using some simple clustering algorithms.

It is particularly important to realize that in this communication system it is not the forms of the transmitted symbols themselves that convey information to the receiver, but rather that the receiver derives its information from recognizing which one out of all the possible symbols was sent during each time interval. In much the same way, the form of an individual symbol printed on this page conveys no information by itself. Information is conveyed to the reader by his recognition of the symbol and whatever meaning he has associated with such recognition.

Since the receiver is essentially a decision making device, all of the possible decisions must be known by the receiver. It is also seen that the only decisions ever made will be those known to exist by the receiver, or those of which the receiver is "aware". For this reason, each received signal will be recognized by the receiver as a representation of one of the symbols it expects the transmitter to send. Each received signal will be recognized as one of the symbols in the receiver symbol set no matter how great the disparity between the transmitter and receiver symbol sets.

It is not particularly important in itself that the receiver and transmitter symbol sets match since the exact form of the symbols has no information content, but the recognition process required of the receiver does require close correspondence between the two sets at least in the features used by the receiver for recognition.

Once the receiver forms a preliminary symbol set and begins to

recognize the incoming signals, the self-evaluating part of the receiver begins to scrutinize the system operation and to propose new symbols, or hypotheses, which lead to more satisfactory recognition of the received signals. Satisfactory in this case, of course, means satisfactory to the receiver.

The communication system receiver proposed uses basic pattern recognition and clustering techniques to recognize the incoming signals and to analyze system performance. Clustering techniques are of particular importance in this investigation because they help the receiver to discover the structure of the data and develop new hypotheses for recognizing as distinct each different symbol in the transmitter symbol set.

Unfortunately, the algorithms used in this part of the study are based on heuristic techniques, that is, techniques which give good results most of the time when applied but which are too complex to permit straightforward mathematical analysis. Because of the algorithms, it is very difficult to show conclusively that the proposed system will ever converge to the optimum state, but it is seen that the learning techniques studied do help bring the system closer to the optimum whenever possible. In general, the learning techniques increase the information flow through the system although it is not certain that the information flow ever reaches the limit imposed by the channel and transmitter.

The rate of information flow through the communication system is used as a measure of the state or amount of learning accomplished by the system. Information flow is perhaps a more meaningful measure of system performance and system learning than many other criteria that

might be proposed since the system's ultimate goal is to transfer knowledge about the choice of symbols made at the transmitter. It is expected that the information flow through the system will increase as the receiver learns and will gradually approach an upper limit determined by the transmitter and channel as the receiver approaches its optimum.

The key to the receiver's operation is that it proposes new recognition schemes at the receiver when the received data indicates to it that those new schemes may aid in improved recognition of the symbols sent by the transmitter. It proposes to receive new symbols when it detects data that may indicate that, for instance, the transmitter is sending two or more different symbols that are not being distinguished by the receiver.

First, the receiver must surmise that two or more of the transmitted symbols are being recognized as identical by the receiver. It does this by examining the past data stored in its memory and detecting inconsistencies in what appears to be noise on the received signals. Exactly what sort of inconsistencies may be detected is determined in large part by the fundamental beliefs of the receiver concerning the transmitter and channel, and its measure of satisfaction with its own operation.

Second, after inconsistencies have been detected, the receiver attempts to improve on this unsatisfactory operation by proposing new symbols using the suspect data itself to model the new symbols. In part, the receiver is automatically endowed with suggestions of how to proceed with the formation of the new symbols because they must be postulated in such a way that the inconsistencies in the data are

resolved.

It may well be that simple adjustments in the receiver's decision scheme will result in satisfactory recognition of the new symbols and, in this case, this is all that is done.

The more interesting situations, though, are those where the data must be reprocessed in order to extract the proper information to permit satisfactory recognition. Generally, this reprocessing results in the supplying of new information to the recognition scheme through a change in the preprocessing of the data. Information in the data which might have been considered irrelevant by the original receiver recognition process may once again be examined in a search for discriminants that lead to satisfactory recognition of the symbols. If such discriminants are found, then they are included permanently in the recognition process for those symbols.

In either case, the way in which the receiver learns the symbols, then, indicates that new symbols are "suggested" by the received data, and that the receiver's dissatisfaction with its own performance starts the learning process.

It should be evident that the receiver must have some bare minimum of knowledge about its own operation and the type of system in which it is to function. Just where this initial knowledge is obtained in a general learning system is a very interesting question but one that this paper does not attempt to answer. It is sufficient to say at this point that the overall response of the system is determined by its basic structure and its concept of how it is to function.

CHAPTER II

THE MODEL

The Communication System Model

The basic communication system model to be utilized in this study is illustrated in Figure 5. The set $\{\bar{x}_i, p_i\}$, $i = 1,2,\ldots k$, represents the k possible symbols sent by the transmitter, each symbol $\bar{x}_i$, an n-dimensional vector and its associated probability p_i. At each signal transmission time, the random vector $\bar{X}$ takes on the value $\bar{x}_i$ with probability p_i. The selection of each $\bar{x}_i$ is independent of all preceeding selections.

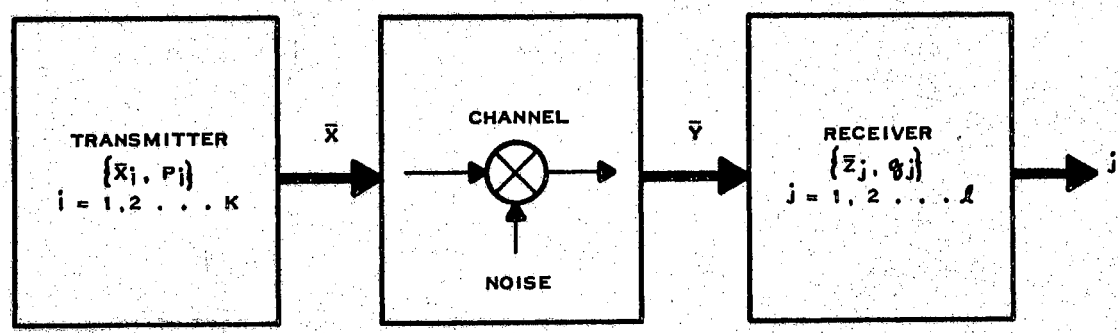

Figure 5. The Basic Model

The channel adds an n-dimensional random noise vector to the transmitted signal $\bar{X}$ to produce the received signal $\bar{Y}$, also an n-dimensional vector. The channel is assumed to be memoryless so the noise vector added to each transmitted signal is independent of the previous signals and noise. The noise vectors are assumed to have a known continuous probability density function $f_{\bar{N}}(\bar{n})$.

For a given transmitted signal $\bar{x}_i$, the random variable $\bar{Y}$ will have density function $f_{\bar{Y}|\bar{X}}(\bar{y}|\bar{x}_i) = f_{\bar{N}}(\bar{y}-\bar{x}_i)$ since the signal and noise are directly additive. Then the unconditional density of $\bar{Y}$ is a mixture of the conditional densities with mixing weights p_i.

$$f_{\bar{Y}}(\bar{y}) = \sum_{i=1}^{k} f_{\bar{Y}|\bar{X}}(\bar{y}|\bar{x}_i)p_i = \sum_{i=1}^{k} f_{\bar{N}}(\bar{y}-\bar{x}_i)p_i \quad .$$

The receiver is a decision-making device which must determine, on the basis of the received signal $\bar{y}$, which one of the $\bar{x}_i$ was sent. At the start of the learning process, the receiver may have only a very limited knowledge about the various $\bar{x}_i$ and their relative frequencies p_i. In order for the receiver to be very effective in making decisions about the symbols being transmitted, it must obtain information about the set $\{\bar{x}_i, p_i\}$ from the received data and then use this information to aid in the decision process.

Considering the receiver as a decision-making device, it is evident that the complete set of possible decisions must be known to the receiver in order that they may be weighed according to some predetermined scheme to arrive at the decision. In effect, this means that each received signal will be identified, or recognized, as one of the symbols which the receiver expects to receive without regard for whether or not

the exact form of the symbol sent by the transmitter matches any of the symbols expected by the receiver.

The set $\{\bar{z}_j, q_j\}$, $j = 1,2,\ldots\ell$, represents the symbols which the receiver expects and their supposed probabilities of transmission, so every signal $\bar{y}$ is recognized as a representation of one of the $\bar{z}_j$'s according to the decision scheme in use at the receiver. In this respect then, the receiver actually receives, or recognizes, only $\bar{z}$'s, and not $\bar{x}$'s, so it is not perhaps technically correct to say that the receiver is able to determine which of the $\bar{x}$'s has been sent. As a convention though, it is to be understood that the receiver is able to identify all the $\bar{x}$'s when it recognizes each different $\bar{x}_i$ as a different $\bar{z}_j$, that is, when each different $\bar{x}_i$ transmitted without noise would be recognized as distinct by the receiver.

Ideally, the receiver is able to recognize the $\bar{x}$'s with the least amount of ambiguity when it expects to receive exactly those symbols and none other. For the receiver, the set $\{\bar{z}_j\}$, $j = 1,2,\ldots\ell$, serves as an estimate of the set $\{\bar{x}_i\}$, $i = 1,2,\ldots k$, and the receiver functions as though the transmitter were actually sending $\bar{z}$'s. Thus, the receiver will be optimum when it has learned the set $\{\bar{x}_i\}$ from the received data and its $\{\bar{z}_j\}$ are chosen so that there is an equivalence between the two sets. That is, there exists a one-to-one relationship between $\{\bar{x}_i\}$ and $\{\bar{z}_j\}$.

Therefore the role of the learning part of the receiver is to examine the received data and use the information it contains, along with any a priori knowledge, to form the set $\{\bar{z}_j, q_j\}$. If the system functions effectively, it is to be expected that the set $\{\bar{z}_j, q_j\}$ will converge to $\{\bar{x}_i, p_i\}$ in some orderly manner as more and more data are

received. The $\{\bar{z}_j, q_j\}$, then, represents the receiver's state of knowledge about the transmitted symbols and their probabilities, $\{\bar{x}_i, p_i\}$.

Pattern Recognition in the Model

Decision-making in the receiver is implemented by the use of pattern recognition techniques. Each received signal $\bar{y}$, an n-dimensional vector, is represented by a point in an n-dimensional Euclidean space at the receiver. The individual $\bar{y}$ vectors are called patterns and the E^n into which they are projected is called the pattern space. The decision scheme is implemented by partitioning the pattern space into regions such that the decision as to which symbol the received signal represents is made by merely noting into which region of the pattern space the pattern $\bar{y}$ falls. The problem of determining the decision process to use is thus seen to be that of finding the proper partition of the pattern space.

Assume for a moment that the receiver has perfect knowledge of the transmitted symbols, i.e., that $\{\bar{z}_j, q_j\} = \{\bar{x}_i, p_i\}$. It is known that the decision scheme yielding lowest error rate under these conditions is the Bayesian decision rule (7). That is, when the signal $\bar{y}_r$ is received, decide that the symbol $\bar{x}_i$ was sent for which $\Pr(\bar{x}_i | \bar{y}_r)$ is maximum. $\Pr(\bar{x}_i | \bar{y}_r)$ means the probability that $\bar{x}_i$ was sent given that $\bar{y}_r$ is the received signal. In practice, Bayes' Theorem may be utilized so that the actual decision is to choose the symbol $\bar{x}_i$ for which $f_{\bar{Y}|\bar{X}}(\bar{y}_r | \bar{x}_i) p_i$ is a maximum. The conditional density function $f_{\bar{Y}|\bar{X}}(\bar{y} | \bar{x}_i)$ is known from the channel characteristics, and p_i is assumed known in this case.

From the pattern recognition point of view, this decision scheme

is equivalent to constructing a partition of the pattern space dividing it up into the k regions where, for each $i = 1,2,\ldots k$, $f_{\overline{Y}|\overline{X}}(\overline{y}|\overline{x}_i)p_i$ is greatest.

The construction of this partition may be simplified considerably if certain assumptions are made about the channel-induced noise. In the problem considered here, it is assumed that the noise is normal and hyper-spherical with zero mean, that is, the noise consists of independent samples drawn from an n-variate normal distribution with zero mean and covariance matrix $\sigma^2 I_n$ where I_n is the n-dimensional identity matrix. Then $f_{\overline{Y}|\overline{X}}(\overline{y}|\overline{x}_i)$ will be a spherical distribution around the point $\overline{x}_i$ since $\overline{y}$ is the sum of the $\overline{x}_i$ vector and the noise vector with zero mean.

Under these assumptions, $f_{\overline{Y}|\overline{X}}(\overline{y}|\overline{x}_i)$ may be expressed as a single-variate function of the square of the Euclidean distance from the point $\overline{y}$ to the point $\overline{x}_i$ without regard for the direction vector between these points.

$$f_{\overline{Y}|\overline{X}}(\overline{y}|\overline{x}_i) = \frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n} \exp\left\{\frac{(\overline{y}-\overline{x}_i)'(\overline{y}-\overline{x}_i)}{2\sigma^2}\right\} .$$

Then $(\overline{y}-\overline{x}_i)$ is distributed as $N(\phi,\sigma^2 I_n)$. So $(\overline{y}-\overline{x}_i)'(\overline{y}-\overline{x}_i) = d^2(\overline{y},\overline{x}_i)$ is distributed as $\chi^2(n)$, a chi-square distribution on n degrees of freedom.

The partitioning surfaces will then be constructed as a set of hyperplanes in the pattern space. For example, suppose there are only two symbols in the transmitter symbol set, $\overline{x}_1$ and $\overline{x}_2$, with respective relative frequencies of transmission p_1 and p_2. The surface partitioning the pattern space into regions with $f_{\overline{Y}|\overline{X}}(\overline{y}|\overline{x}_1)p_1 \geq f_{\overline{Y}|\overline{X}}(\overline{y}|\overline{x}_2)p_2$ in

one of the regions and with the inequality reversed in the other region is a hyperplane perpendicular to the line segment joining $\bar{x}_1$ and $\bar{x}_2$ and passing through the point on that line where $f_{\bar{Y}|\bar{X}}(\bar{y}|\bar{x}_1)p_1 = f_{\bar{Y}|\bar{X}}(\bar{y}|\bar{x}_2)p_2$.

A one-dimensional example of this is shown in Figure 6 where the hyperplane is merely a point on the y-axis. When the signal y is received, the decision is made by comparing the value of y with d. If $y \geq d$, decide x_2 was sent, otherwise decide x_1 was sent. The point $y = d$ is the solution to the equation $f_{\bar{Y}|\bar{X}}(y|x_1)p_1 = f_{\bar{Y}|\bar{X}}(y|x_2)p_2$. This decision scheme will yield the lowest decision error rate for the system described.

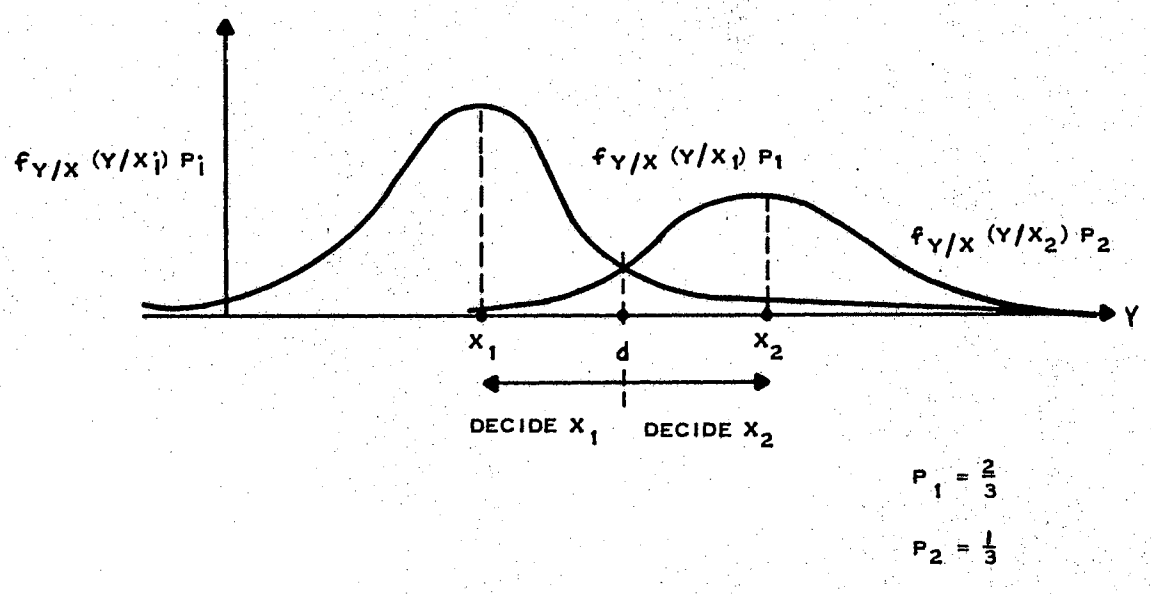

Figure 6. Bayesian Decision Scheme in one Dimension

The Receiver Decision Scheme

In the actual learning system, of course, the receiver has no such perfect knowledge of the transmitted symbols and their probabilities upon which to base the decision process. Under these circumstances, it appears that the receiver can only utilize the best information available to it, that being the $\{\bar{z}_j, q_j\}$, $j = 1,2,\ldots\ell$, which it uses as its estimate of the transmitted symbols and their probabilities.

The receiver, as explained previously, must operate as if the transmitter were actually sending symbols from $\{\bar{z}_j\}$, $j = 1,2,\ldots\ell$, and so the receiver decision scheme operates as though this were the case. The receiver is assumed to have complete knowledge of the channel noise and so it is able to consider probabilities such as $\Pr(\bar{z}_j \text{ received} \mid \bar{z}_i \text{ sent})$. An outside observer, of course, knows that the symbol $\bar{z}_i$ is not actually sent by the transmitter (unless it happens to coincide with one of the $\bar{x}$'s), and so such a probability is fictitious. However, the receiver "thinks" that the symbol $\bar{z}_i$ is being sent by the transmitter with relative frequency q_i, so such a probability has a definite and logical meaning to the receiver even though, in fact, the event may never occur.

As a notational convention, the letter z will represent symbols received, or recognized, by the receiver, and z' will be used to indicate symbols which the receiver supposes are present among the set at the transmitter. Thus, $\Pr(\bar{z}_j \text{ received} \mid \bar{z}_i \text{ sent})$ is written $\Pr(\bar{z}_j \mid \bar{z}'_i)$. The letter x indicates the symbols actually transmitted, and y the resulting signals presented to the receiver.

In view of this, the decision scheme used by the receiver is: upon

receipt of the signal $\bar{y}_r$, decide that the symbol $\bar{z}_i$ was sent for which $Pr(\bar{z}_i' | \bar{y}_r)$ is a maximum. Bayes' Theorem is again utilized so that the actual decision rule is: upon receipt of a signal $\bar{y}_r$, decide that the symbol $\bar{z}_i$ was sent for which $f_{\bar{Y}|\bar{Z}}(\bar{y}_r | \bar{z}_i')q_i$ is a maximum.

It would be a surprising coincidence if the decision rule based upon the $\{\bar{z}_j, q_j\}$, $j = 1,2,\ldots\ell$, were optimum for recognizing all the transmitted symbols. However, as the receiver in learning the transmitted symbols brings the $\{\bar{z}_j, q_j\}$ closer and closer into correspondence with the $\{\bar{x}_i, p_i\}$, and the decision rule changes accordingly, the receiver's ability to recognize the transmitted symbols should improve until it approaches the limit imposed by the channel noise.

The Model Receiver

A block diagram of a complete adaptive receiver is shown in Figure 7. Briefly, the received signals are preprocessed in the Measurement Selector and then presented to the Categorizer which performs the decision process, symbol recognition, with the aid of the information contained in $\{\bar{z}_j, q_j\}$, $j = 1,2,\ldots\ell$. The received signals, along with the output of the decision process, are stored in the Memory for later reference by the Performance Evaluator and Symbol Generator which adjusts the receiver symbol set as learning progresses.

The Measurement Selector is perhaps the simplest possible type of preprocessor. The n-dimensional vector $\bar{y}$ at the input is reduced to an n_j-dimensional vector, $0 < n_j \leq n$, at the output by simply disregarding the last $n-n_j$ components of the vector $\bar{y}$. That is, the first n_j components of $\bar{y}$ are passed unchanged through the Measurement Selector and outputted as an n_j-dimensional vector. This vector will be denoted by

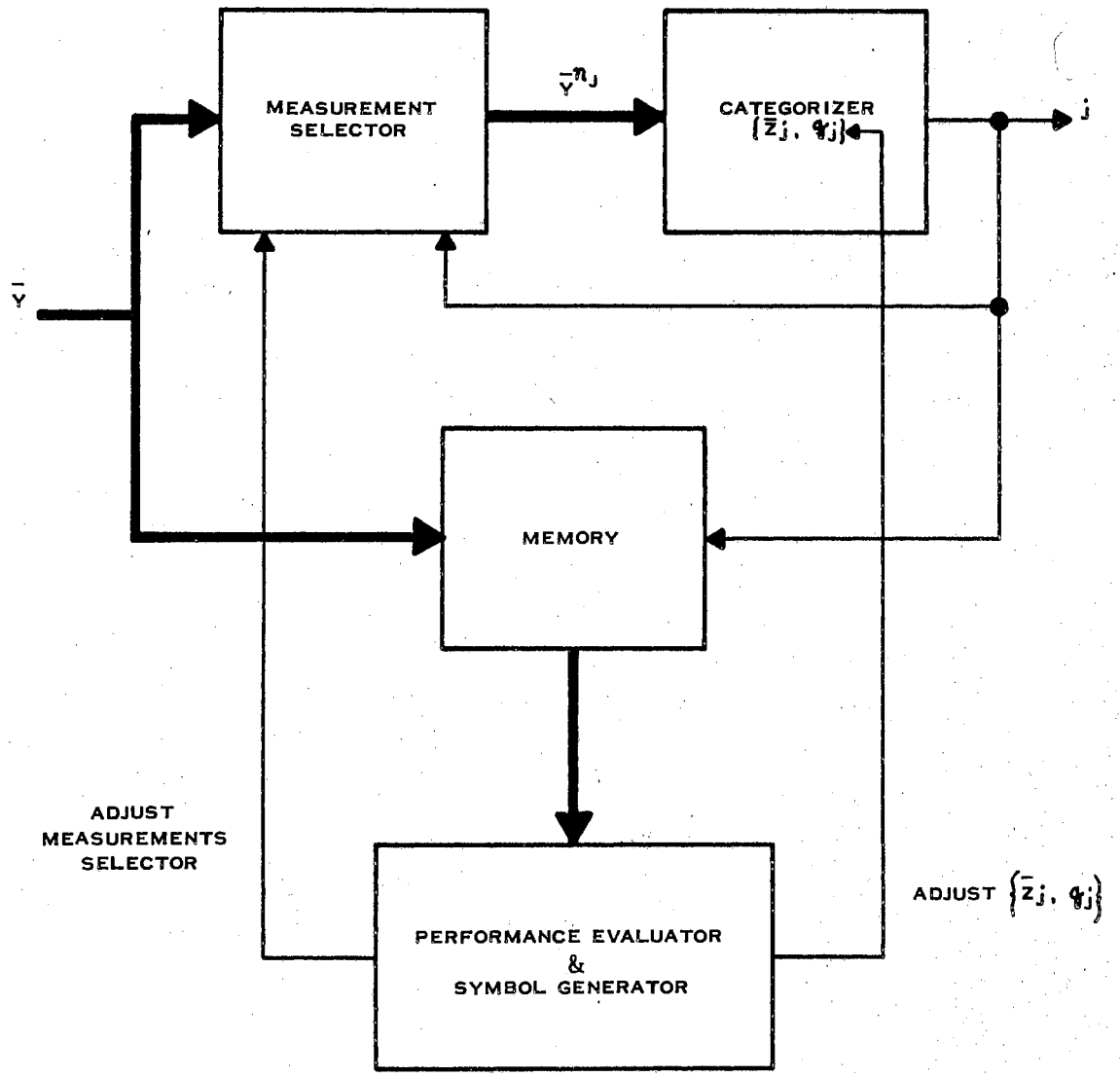

Figure 7. The Adaptive Receiver Model

$\bar{y}^{n_j}$ when it is necessary to indicate the dimensionality of the vector. Any information contained in the last $n-n_j$ components of $\bar{y}$ does not contribute to the decision process.

The number n_j is the number of components of $\bar{y}$ which are utilized at a given stage of the learning process for recognition of y as a representation of the jth symbol, $\bar{z}_j$. The number of components is, in general, to be held to the minimum that will produce acceptable results in the recognition process. This means in effect that the information used in the decision is determined by the decision itself, a seemingly circular situation.

In the model system this difficulty is circumvented by performing a series of decisions, each requiring as much or more information than all previous decisions until the process terminates when one and only one of the $\bar{z}_j$ is left for which $\bar{y}_r$ may be a representation. The overall recognition process may be represented by a tree diagram like Figure 8 where each succeeding decision is made using more and more components of the received signal vector $\bar{y}_r$. The number of dimensions to be considered in each decision is determined by the outcome of the immediately preceeding decision in the tree.

The pattern space in which the decision surfaces are constructed needs only n_j dimensions. As the categorizer performs the sequence of decisions at each stage of the decision tree, the dimensionality of the pattern space and decision surfaces thus increases just enough to perform the perceived task at hand.

At this point it should be evident that the vector representations of the recognized symbols, $\{\bar{z}_j\}$, need be vectors each containing only n_j elements since the decisions made with the use of these vector models

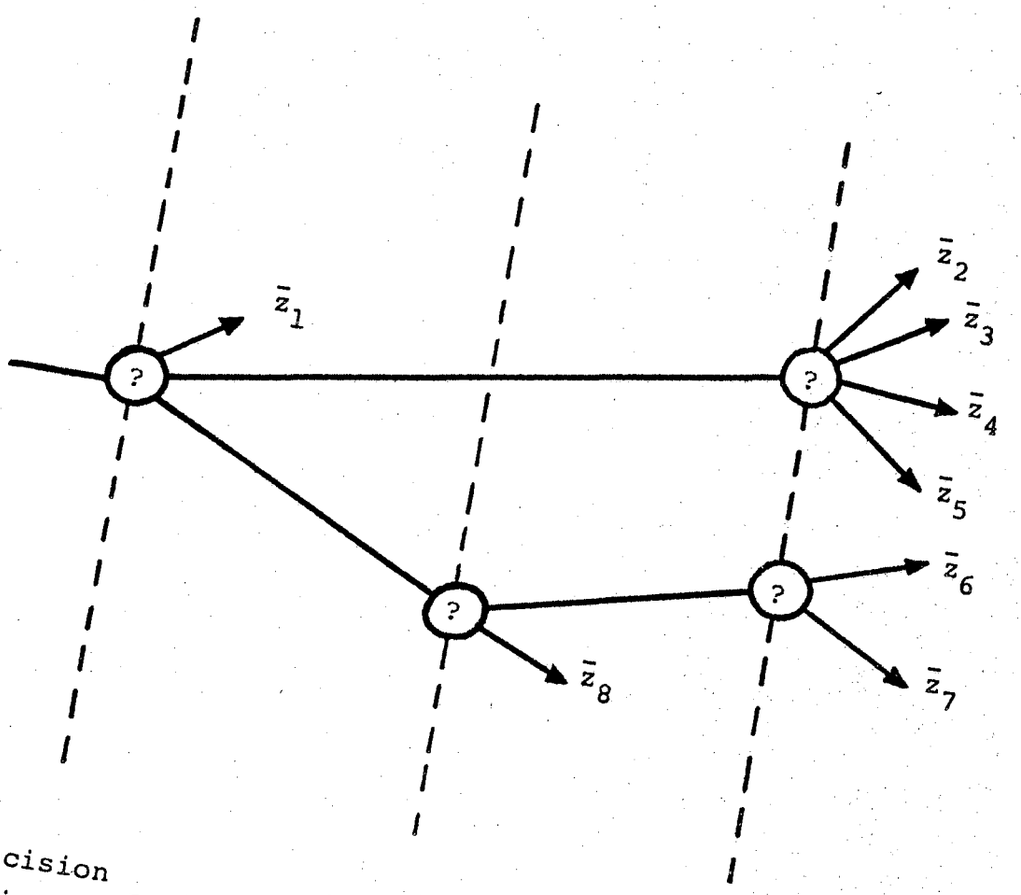

Figure 8. Example of Decision Sequence

of the received symbols are made on the basis of only n_j components of the received signals. Thus, the receiver stores a description of the symbols it expects to receive only to the extent necessary to recognize those symbols as distinct from each other.

The function of the Performance Evaluator and Symbol Generator is to periodically examine the Memory, which contains the past data and performance, and to propose new symbols with the corresponding recognition scheme which will cause the system operation to become more "satisfactory" according to the criteria built into the Performance Evaluator section.

The proposal of new symbols is in the form of modifications to the $\{\bar{z}_j, q_j\}$, $j = 1,2,\ldots\ell$, through the addition or deletion of $\bar{z}$'s in the set and a change in the number ℓ, or a modification of the $\bar{z}$'s without changing ℓ, perhaps by changing the dimension of some symbol representations and by adjusting some component values. The exact manner in which these adjustments are made can be more easily explained after a discussion of some general clustering techniques.

Clustering Techniques and Decomposing Mixtures of Distributions

Clustering, or similarity grouping, is closely related to the statistical problem of the decomposition of mixtures of distribution functions. This study is concerned with the use of clustering techniques in a system that learns without a teacher, and particularly the use of heuristic clustering methods in the learning process. It may, however, be informative to briefly consider the statistical problem of decomposing mixtures of distribution functions and the relationship between this problem and clustering methods.

In the general problem of decomposing a mixture of distribution functions, it is desired to identify various parameters of the mixing distributions and their proportions in the mixture using information derived from a sample drawn from the mixture. That is, when samples are drawn from the density of the mixture $f(\bar{x}) = \sum_{i=1}^{k} p_i f_i(\bar{x})$ and the individual densities $f_i(\bar{x})$ are known except for some parameters, the problem is to find estimates for the p_i and any unknown parameters in the $f_i(\bar{x})$.

One method of solution involves solving a set of simultaneous equations of the moments of the mixture and of the sample. To be sure, this method is quite successful in a number of instances. It is, however, necessary to know the number of mixing distributions and, when this number is large, the moment equations get rather complicated. In addition, if any real accuracy is necessary in the estimates, the sample size must be quite large (8).

A more efficient method based on the likelihood function is apt to be even more complex except in special cases (8).

Certainly these solutions to the problem are statistically satisfying, but the complexity of the methods severely limits their usefulness in practical situations. Clustering techniques, on the other hand, are based more on a straightforward heuristic approach to the solutions and, in general, cannot be analyzed in simple statistical terms (9).

The clustering problem is somewhat the same as that just described. One is presented with a sample drawn from a mixture of distributions and it is desired to group the samples in such a way that the individual clusters as nearly as possible have all of their members arising from the same single distribution. The mixing proportions can then be

estimated from the number of samples in each cluster, and any unknown parameters in the mixing distributions may be estimated by considering each cluster on an individual basis.

In the communication system model, the input to the Receiver is the random variable $\bar{Y}$, a sample drawn from the mixture density $f_{\bar{Y}}(\bar{y}) = \sum_{i=1}^{k} p_i f_{\bar{Y}|\bar{X}}(\bar{y}|\bar{x}_i)$. The problem is to estimate the $\{p_i\}$ and the means of the conditional densities, the $\{\bar{x}_i\}$. This may be done by identifying the sub-sets arising from each of the conditional densities.

If the $\bar{x}_i$ are sufficiently separated from each other in terms of the standard deviation of the channel noise, the $\bar{y}$ samples will be distributed in well separated clusters of points in the pattern space. The shape of each cluster will be similar to the channel noise distribution. This knowledge may be used to identify the individual clusters through the use of some relatively simple algorithms such as that described below (9).

Let the set of N sample patterns be denoted by $\{\bar{y}_r\}$, $r = 1,2,\ldots N$, and let $D(\bar{a},\bar{b})$ be a metric defined on all points, or patterns, in the pattern space which measures the distance between two such points, $\bar{a}$ and $\bar{b}$. Let Z_j, $j = 1,2,\ldots \ell$, be the set of points classified as members of the jth cluster, with cluster center $\bar{z}_j$, usually the mean of Z_j. Every sample $\bar{y}_r$ is then assigned to one and only one of the clusters by the rule: $\bar{y}_r \epsilon Z_j$ iff $D(\bar{y}_r,\bar{z}_j) = \min_s D(\bar{y}_r,\bar{z}_s)$. The sample point $\bar{y}_r$ is thus assigned to the cluster to the center of which it is closest in the sense of the metric D. Ties may be settled by any convenient method, either random or deterministic.

In general, the cluster centers $\bar{z}_j$ are only estimates of the means of the mixing distributions since the true means are unknown to the

receiver. After all the samples are classified, then, the cluster centers $\bar{z}_j$ are redefined as the means of the sets Z_j, $j = 1,2,\ldots\ell$. Then the classification step is repeated, the cluster centers redefined, etc., until the procedure converges.

If $\bar{z}_j^t$ and Z_j^t are the jth cluster center and the set of samples classified into the jth cluster, respectively, on the tth iteration, the clustering algorithm may be stated as:

1) assign every sample $\bar{y}_r$ to one and only one cluster so that $\bar{y}_r \epsilon Z_j^t$ iff $D(\bar{y}_r, \bar{z}_j^t) = \min_s D(\bar{y}_r, \bar{z}_s^t)$ (1a) except that ties may be settled arbitrarily;

2) redefine the cluster centers as

$$\bar{z}_j^{t+1} \ni \sum_{\bar{y}_r \epsilon Z_j^t} D(\bar{y}_r, \bar{z}_j^{t+1}) = \min_{\bar{y}} \sum_{\bar{y}_r \epsilon Z_j^t} D(\bar{y}_r, \bar{y}) \quad . \qquad (1b)$$

It can be shown that the algorithm converges since the sum of the distances from the points in the clusters to their respective cluster centers is non-increasing in both steps. Consider the case of a fixed set of sample patterns $\{\bar{y}_r\}$, $r = 1,2,\ldots N$, and a fixed number of clusters, say ℓ. Then Z_j^t, $j = 1,2,\ldots\ell$, denotes the subset of sample patterns classified into the jth cluster on the tth iteration of the algorithm. Note that $\cup\, Z_j^t = \{\bar{y}_r\}$ and $Z_j^t \cap Z_k^t = \emptyset$ $j, k = 1,2,\ldots\ell$, $j \neq k$. As before, let $\bar{z}_j^t$, $j = 1,2,\ldots\ell$, denote the cluster centers. Now define the sum of the distances from the cluster center to all points in that cluster on the tth iteration by

$$c_j^t = \sum_{\bar{y}_r \epsilon Z_j^t} D(\bar{y}_r, \bar{z}_j^t) \qquad j = 1,2,\ldots\ell$$

and let

$$C^t = \sum_{j=1}^{\ell} c_j^t = \sum_{j=1}^{\ell} \sum_{\bar{y}_r \epsilon Z_j^t} D(\bar{y}_r, \bar{z}_j^t)$$

denote the total of all the distances from the sample patterns to their respective cluster centers on the tth iteration. In addition to the above, define the distances $c_j'^t = \sum_{\bar{y}_r \epsilon Z_j^t} D(\bar{y}_r, \bar{z}_j^{t+1})$ and $C'^t = \sum_{j=1}^{\ell} c_j'^t$.

The clustering algorithm is started using a set of cluster centers $\bar{z}_j^1$ picked by any method, perhaps at random. The first step of the algorithm assigns each sample pattern to one of the clusters and permits computation of the distance sums c_j^1, $j = 1,2,\ldots\ell$, and C^1.

Application of the second step of the algorithm defines new cluster centers $\bar{z}_j^2$ such that

$$\sum_{\bar{y}_r \epsilon Z_j^1} D(\bar{y}_r, \bar{z}_j^2) = \min_{\bar{y}} \sum_{\bar{y}_r \epsilon Z_j^1} D(\bar{y}_r, \bar{y}) \leq \sum_{\bar{y}_r \epsilon Z_j^1} D(\bar{y}_r, \bar{z}_j^1) = c_j^1 \qquad j=1,2,\ldots\ell.$$

But note that we have defined $c_j'^1 = \sum_{\bar{y}_r \epsilon Z_j^1} D(\bar{y}_r, \bar{z}_j^2)$ so we have the result that $c_j'^1 \leq c_j^1$ $j = 1,2,\ldots\ell$. It easily follows that $C'^1 \leq C^1$.

At this point the overall sum of distances is equal to

$$C'^1 = \sum_{j=1}^{\ell} \sum_{\bar{y}_r \epsilon Z_j^1} D(\bar{y}_r, \bar{z}_j^2)$$

and because of the mutually exclusive and exhaustive properties of the

sets Z_j there are exactly N distance terms in the sum, one for each $\bar{y}_r$.

Application of the first step of the clustering algorithm redefines the categories. The new categories are assigned in such a way that for each $\bar{y}_r$

$$D(\bar{y}_r, \bar{z}_j^2) = \min_s D(\bar{y}_r, \bar{z}_s^2)$$

and so the individual distances are minimized by the category assignments. Therefore

$$\sum_{j=1}^{\ell} \sum_{\bar{y}_r \in Z_j^2} D(\bar{y}_r, \bar{z}_j^2) \leq \sum_{j=1}^{\ell} \sum_{\bar{y}_r \in Z_j^1} D(\bar{y}_r, \bar{z}_j^2)$$

where the equality holds when no changes take place in the classifications.

Note that

$$C^2 = \sum_{j=1}^{\ell} \sum_{\bar{y}_r \in Z_j^2} D(\bar{y}_r, \bar{z}_j^2) \leq \sum_{j=1}^{\ell} \sum_{\bar{y}_r \in Z_j^1} D(\bar{y}_r, \bar{z}_j^2) = C'^1 \leq C^1$$

or

$$C^2 \leq C^1 \quad .$$

Exactly the same argument shows that on the t<u>th</u> iteration, $C^{t+1} \leq C^t$. Also note that $C^t \geq 0$ because all the distances are non-negative. Then the sequence C^t is a non-increasing sequence which is bounded below and therefore is a convergent sequence.

The function C^t is a measure of the compactness of the clusters so

apparently the clusters tend to become more compact in the sense of the metric D with each iteration.

The more interesting question of the limit to which the clustering algorithm converges remains unanswered. It would be very convenient if convergence to the most compact grouping were assured, but it is easily shown by a counter-example that this is not always the case (10). The clustering algorithm may converge to a local minimum. The convergence properties of the algorithm are dependent upon the distribution of the sample patterns, the number of clusters used in the algorithm, and the starting points for the cluster centers. These same factors determine the number of iterations required for convergences.

One might reasonably expect the algorithm to find the most compact grouping with high probability when the distribution of the sample patterns is in well-separated, unimodal clusters and, most important, when the supposed number of clusters used in the algorithm coincides with the true number of clusters in the sample patterns. If the numbers of real and supposed clusters are not the same, the algorithm will nevertheless group the samples into the supposed number of clusters whether or not the resulting clusters are very compact according to some absolute scale.

It is obvious that the number of clusters must either be known a priori or obtained somehow from the data. In the statistical separating methods mentioned earlier, it is necessary to know the number of mixing distributions, or clusters, before applying the methods. In the algorithmic procedure just discussed, it is equally important to have some method of arriving at the correct number of clusters.

Clustering Techniques in the Receiver

In the model receiver, the received signal vectors are projected as points into the pattern space and then arranged into similar groups using the clustering algorithm discussed previously. The metric D used in the clustering algorithm is taken to be the reciprocal of the likelihood function used as the receiver decision scheme. That is, the sample $\bar{y}_r$ is classified into the jth cluster for which $f_{\bar{Y}|\bar{Z}}(\bar{y}_r|\bar{z}_j)q_j$, $j = 1,2,\ldots\ell$, is a maximum.

Heuristic methods, along with knowledge of the channel noise, are used in the model to find the correct number of clusters for use with the clustering algorithm. New cluster centers are proposed to accept samples that are far away from all the previously assumed clusters, and two or more clusters are combined into one if their centers closely approach each other. By this means it is expected that the system will be able to determine the correct number of clusters as it learns from the received data. A qualitative idea of the values to assign to "far away" and "close" can be determined from knowledge of the channel noise along with some considerations about the type of environment in which the receiver is expected to operate.

As a preliminary assumption, it is proposed that the symbols sent by the transmitter be readily identifiable in spite of the channel noise when the proper features are used in the discrimination process. By "readily identifiable" is meant that the optimum decision scheme would recognize the symbols with only, say, about 1% or less misclassification. If the actual error rate is higher, it simply means either that the features selected are not the best, or that the decision scheme is not utilizing the information in an efficient manner.

In view of this, it is evident that the point clusters formed by the received signals in the pattern space may be easily separated when the proper measurements are made on the signals. Therefore, for a given set of features upon which to perform classification, it will initially be desirable to identify as distinct only those clusters which are reasonably well separated from each other. The problem of what to do about clusters which are too close for separate identification will be taken up later.

Consider first the question of how far from all existing clusters a new sample point should be in order to cause a new cluster center to be generated by the system. In the context of the communication system problem, this is equivalent to asking how much different from all existing representations of the symbols a received signal should be before the signal should be taken to represent a new symbol.

Suppose that there have been m received patterns $\{\bar{y}_r\}$, $r = 1,2,...m$, classified into the same cluster, Z_a, and that, in fact, all m patterns were generated by the same symbol, $\bar{x}_a$, at the transmitter. If the channel noise added to the transmitted signal has a normal distribution with mean vector $\emptyset$ and covariance matrix $\sigma^2 I_n$, then the m received signals may be considered to have been drawn from a $N(\bar{x}_a, \sigma^2 I_n)$ distribution. If the cluster center, $\bar{z}_a$, is taken to be the mean of the received signals, then $\bar{z}_a = \frac{1}{m} \sum_{r=1}^{m} \bar{y}_r$, and $\bar{z}_a$ will also be a normal random vector with mean $\bar{x}_a$ and covariance matrix $\frac{\sigma^2}{m} I_n$.

When the m+1<u>th</u> signal, $\bar{y}_{m+1}$, is received, the system is required to make a decision as to whether $\bar{y}_{m+1}$ should be classified into Z_a (assuming that $\bar{y}_{m+1}$ is within the existing decision region for Z_a), or whether a new cluster should be proposed to accept $\bar{y}_{m+1}$. This question

may be rephrased in terms of testing the hypothesis that $\bar{y}_{m+1}$ is from the same $N(\bar{x}_a, \sigma^2 I_n)$ distribution as the previously received signals. Consider $H_0: \bar{y}_{m+1}$ is from the same $N(\bar{x}_a, \sigma^2 I_n)$ distribution as $\{\bar{y}_r\}$, $r = 1,2,\ldots m$, and $H_1: \bar{y}_{m+1}$ is from a distribution with different mean value but same covariance matrix, i.e., that $\bar{y}_{m+1}$ resulted from the transmitter sending some new symbol other than $\bar{x}_a$. Under H_0, the difference $\bar{y}_{m+1} - \bar{z}_a$ is distributed as $N(\emptyset, \frac{m+1}{m} \sigma^2 I_n)$ and the square of the Euclidean distance between $\bar{y}_{m+1}$ and $\bar{z}_a$ is distributed as $\frac{m+1}{m} \sigma^2 \chi^2(n)$. That is, $md^2(\bar{y}_{m+1}, \bar{z}_a)/(m+1)\sigma^2$ is distributed as $\chi^2(n)$.

The critical region for this test will be taken to be the tail of the distribution so that the significance level of the test is $\alpha = \int_{t_\alpha}^{\infty} \chi^2(n) d\chi^2$. If $md^2(\bar{y}_{m+1}, \bar{z}_a)/(m+1)\sigma^2 > t_\alpha$, then the hypothesis that $\bar{y}_{m+1}$ is from the same distribution as the samples used to form $\bar{z}_a$ is rejected and $\bar{y}_{m+1}$ is used as the start of another cluster. For example, if n, the dimensionality of the vectors, is 5, and m, the number of samples, is 10, then a new cluster will be started when the eleventh sample is such that

$$\frac{10\ d^2(\bar{y}_{11}, \bar{z}_a)}{11\ \sigma^2} > 15.086 \text{ or } d(\bar{y}_{11}, \bar{z}_a) > 4.15\ \sigma$$

at the 1% significance level.

Now consider the power of the test when $\bar{y}_{m+1}$ is a sample drawn from a distribution with mean $k\sigma$ away from the mean of the distribution from which the samples in Z_a were drawn. Under H_1, $md^2(\bar{y}_{m+1}, \bar{z}_a)/(m+1)\sigma^2$ will be distributed as $\chi'^2(n,\lambda)$ where

$$\lambda = \frac{k^2 m}{2(m+1)} \quad .$$

The power of the test $= \int_{t_\alpha}^{\infty} \chi'^2(n,\lambda)d\chi'^2$. For example, if $k = 5$, then $\lambda = \frac{25(10)}{2(11)} = 11.35$ and the power of the test is approximately 0.51 at the 1% significance level.

Next, consider the question of when to combine two clusters into one larger cluster. One might test the hypothesis that two clusters actually have all their samples arising from the transmission of a single symbol, but the basic assumption that all the symbols are readily identifiable when the right features are examined, and that decision schemes with high error probabilities are not to be considered, suggests a more meaningful criterion for combining clusters. It may be more useful to combine clusters when it appears that there is no decision scheme available for distinguishing separate symbols with suitably low probability of error, say, on the order of 1% or less.

Suppose two symbols resulting from the transmission of $\bar{x}_a$ and $\bar{x}_b$ are recognized at the receiver on the basis of n features. If these symbols are to be recognized with only 1% or less probability of error and both symbols are equally likely, it is necessary for the vector representatives of the n features of $\bar{x}_a$ and $\bar{x}_b$ used in the recognition process to be at least 4.66σ apart in the n-dimensional pattern space. It might be noted here that the "recognizability" of the symbols is independent of the dimensionality of the pattern space. If the statistics of the noise in each dimension and the Euclidean distance between the features observed are held constant while changing the dimensionality, the symbols may be recognized with a constant probability of error independent of n. With this in mind, the representations of any two symbols at the receiver, $\bar{z}_a$ and $\bar{z}_b$, may be used in a test to decide

whether or not $\bar{z}_a$ and $\bar{z}_b$ are indeed derived from samples of two transmitted symbols that are at least 4.66σ apart in the n-dimensional vector of features used by the recognition scheme.

Suppose that the system has formed two clusters, Z_a and Z_b, with centers $\bar{z}_a$ and $\bar{z}_b$ on the basis of m_a and m_b samples, respectively, being classified into each category. Further suppose that all the samples in cluster Z_a were actually generated by the transmission of symbol $\bar{x}_a$, and likewise for cluster Z_b and transmitted symbol $\bar{x}_b$. Then the cluster centers $\bar{z}_a$ and $\bar{z}_b$ are n-dimensional vector estimates of the n features of $\bar{x}_a$ and $\bar{x}_b$. These estimates, $\bar{z}_a$ and $\bar{z}_b$, are random variables, of course, with distributions $N(\bar{x}_a, \frac{\sigma^2}{m_a} I_n)$ and $N(\bar{x}_b, \frac{\sigma^2}{m_b} I_n)$, respectively.

Using these estimates with known distributions, it is a simple matter to test the hypotheses H_0:$\bar{x}_a$ and $\bar{x}_b$ are $k\sigma$ or more apart in the n features represented by $\bar{z}_a$ and $\bar{z}_b$, and H_1:$\bar{x}_a$ and $\bar{x}_b$ are less than $k\sigma$ apart in those n features. The vector difference $\bar{z}_a - \bar{z}_b$ will be distributed as $N(\bar{x}_a - \bar{x}_b, \frac{m_a+m_b}{m_a m_b} \sigma^2 I_n)$. Therefore, similar to the previous test, it is seen that under H_0 where $d^2(\bar{x}_a, \bar{x}_b) = k^2\sigma^2$, the test statistic $\frac{m_a m_b d^2(\bar{z}_a, \bar{z}_b)}{(m_a+m_b)\sigma^2}$ is distributed as $\chi'^2(n,\lambda)$ where $\lambda = \frac{k^2 m_a m_b}{2(m_a+m_b)}$.

The critical region will be taken to be the left-hand end of the distribution so H_0 will be rejected at significance level α when

$$\frac{m_a m_b d^2(\bar{z}_a, \bar{z}_b)}{(m_a+m_b)\sigma^2} < t_\alpha$$

where

$$\alpha = \int_0^{t_\alpha} \chi'^2(n,\lambda) d\chi'^2$$

and

$$\lambda = \frac{k^2 m_a m_b}{2(m_a+m_b)} \quad .$$

For example, at the 1% significance level with $n = 5$, $m_a = m_b = 10$, and $k = 4.66$, the clusters Z_a and Z_b will be combined when $\frac{100\ d^2(\bar{z}_a, \bar{z}_b)}{20\ \sigma^2} < 29.9$. When $k = 2.33$, half the previous value, the power of the test is about 0.91.

In general, things cannot be expected to be quite as simple as is assumed in the examples because, for instance, there is no assurance that all the samples classified into Z_a and used to form the cluster center $\bar{z}_a$ actually did arise from the transmission of symbol $\bar{x}_a$. However, the examples do indicate some general guidelines for finding decision points that will give useful results when used in the learning system. For the situation cited in the examples it seems evident that the learning system will perform its task in a reasonably efficient manner if a new cluster is started to accept a received signal more than about 4.1σ away from all existing cluster centers or if two clusters are combined when their centers become less than about 2.4σ apart. Of course these decision points change with the number of samples received. The learning receiver may calculate the values as they are required.

The growth of a set of reasonably compact and well separated clusters is essential to the initial progress of learning in the model. It is expected that the cluster starting and combining tests will produce such a set when they are reasonably well "balanced", that is, the rate of generation of new clusters by the one test equals the rate of extinction of clusters by the other test.

It is easy to see that the new cluster starting test will, with probability one, generate new clusters whether they are valid or not as the number of received signals grows. Suppose that there is only one

transmitted symbol so all the received signals are drawn from a single distribution. Under the assumption that the signals are normally distributed, the distances between the mean and some of the signals will exceed any threshold if the sample is large enough. It would seem that there is the possibility that the number of clusters could grow without bound.

The limiting behavior of the symbol combining test is not so easily analyzed and it has not been proven that the rate of cluster extinction equals the cluster growth rate. Nevertheless, it is intuitively evident that the cluster extinction rate must grow with the number of clusters generated by the cluster starting test from a single compact distribution. As more and more samples are drawn from the single distribution, the clusters will tend to become grouped near the mode of the distribution (for the unimodal distribution considered here) and will eventually be close enough to be combined by the symbol combining test. This should take place at a rate equal to the generation of spurious clusters. This supposition is well supported by the experimental results presented later.

Hypothesis Generation in the Model

By following the previously explained procedures, it is expected that the learning receiver will develop a decision scheme which separates the received signals into fairly well separated clusters based on the minimum number of features selected for classification by the system. It is understood that the features selected by the system may not contain sufficient information to permit the proper classification, or recognition, of each different transmitted symbol as distinct from

all the others. Therefore, it is quite possible that the decision scheme found by the above procedure will recognize two or more transmitted symbols as identical. On the other hand, it would be rather improbable that the system would consistently recognize the same transmitted symbol as two or more different symbols at the receiver since this is equivalent to dividing up one cluster of points in the pattern space into several distinct groups, and such a grouping satisfying the symbol combining test is unlikely in samples drawn at random from one cluster of points, at least when the number of samples is large.

Thus, the first objective of the learning receiver at this point is to determine which clusters of points represent more than one transmitted symbol and then to find ways of dissecting the multiple-symbol clusters so that each individual symbol is recognized distinctly.

First, it is necessary to detect the clusters of points which might represent more than one symbol. One would expect the sample points in a cluster arising from a single transmitted symbol to have the same distribution as the channel noise, and it would be possible to test each cluster of points to decide whether or not that cluster is composed of samples drawn from a distribution of the same type as the channel noise. Any significant discrepancy would suggest that the cluster is composed of samples generated by the transmission of more than one symbol. Since the channel noise distribution is assumed known, this approach is entirely feasible; however, a more general approach is used in the model receiver with a view toward developing methods not dependent upon a complete knowledge of the channel noise.

The method of detecting multiple-symbol clusters which is proposed for use by the system is based upon the fact that one would expect all

clusters of points arising from the transmission of single symbols to have approximately the same shape, or distribution, that shape being determined by the channel noise. Also, one would also expect that the most compact clusters are, in fact, single-symbol clusters. Thus, the system is set up to operate on the premise that compact clusters represent single symbols and that large, spread-out clusters are probably combinations of single-symbol clusters whose symbols are mistakenly being recognized as identical.

The model system finds the most compact cluster with at least an average number of points and takes the distribution of the points in that cluster to be typical of the channel noise distribution. Then a comparison is made between this most compact cluster and the next most compact cluster to decide whether or not the two clusters could have been drawn from distributions identical except for the means. The test used for this purpose was derived by Bishop (11) and is suggested by the fact that for the normal noise distribution case the cluster shape is determined by the sample covariance matrix. The hypothesis H_0: both clusters were drawn from distributions with identical covariance matricies, is tested against H_1: the two clusters were drawn from distributions with different covariance matricies. The test compares the determinants of the two sample covariance matricies. The determinant of the covariance matrix is a measure of the compactness of the cluster and any significant difference in the determinant is taken to indicate a difference in the matricies and a corresponding difference in cluster shapes. If the test reveals no significant difference in the underlying distributions, then it is presumed that both clusters represent single symbols sent by the transmitter. On the other hand, if it appears that

there is a significant difference in the cluster distributions, the larger cluster is suspected of representing more than one transmitter symbol and methods must be found to refine the decision scheme creating that cluster so the separate symbols may be recognized as distinct.

For purposes of this discussion, assume that the set of clusters $\{Z_j\}$, $j = 1,2,\ldots\ell$, has been ordered so that the subscript denotes the relative compactness of each cluster as measured by the determinant of the sample covariance matrix, with Z_1 being the most compact and Z_ℓ the least compact. Then Z_1 is assumed by the model to represent a typical single-symbol cluster, and Z_1 and Z_2 are tested to decide whether or not the covariance matricies of their respective distributions are similar. If the test reveals no significant difference, Z_1, Z_2, and Z_3 are all used in the same test to check for significant differences in their covariance matricies. As long as the test reveals no significant differences in the covariance matricies, more and more of the clusters, in order of their compactness, are included in the test until a difference is detected or else all the clusters have passed the test.

Suppose that the test indicates a significant difference in cluster, Z_j, $1 < j \leq \ell$. An attempt is first made to distinguish the individual clusters in Z_j on the basis of the symbol features then in use by the system. It is possible that under certain conditions the clustering algorithm may cause two or more readily separable clusters to be grouped into one large cluster; in such a situation no new information in the form of additional symbol features should be required to perform satisfactory separation of the individual clusters.

If two readily separable clusters are indeed grouped as one, the best way to divide the two clusters is by a plane perpendicular to the

direction of greatest dispersion of the overall cluster. The direction of greatest dispersion is the direction of the largest eigenvector associated with the covariance matrix of the overall cluster. If there were an equal number of samples in each of two component clusters, the dividing plane should pass through the composite mean.

In the general situation, things are not quite so straight-forward since, in principle at least, there may be any number of component clusters with different numbers of samples in each making up the large cluster. Nevertheless, it seems that a good way to break up the anamolous cluster is to start with one dividing plane perpendicular to the largest eigenvector and to search along that vector until an "open space" is found which would seem to be a natural location for the dividing plane. This is done in the model system by arbitrarily dividing the cluster into 20 "bins" using 19 equally-spaced planes perpendicular to the largest eigenvector. The bin with the largest number of sample points is found, representing one mode of the cluster. Another mode is found some distance away and a dividing plane is passed through the bin between the two modes which has the smallest number of sample points.

The clusters formed on each side of the dividing plane may then be refined by applying the clustering algorithm with the number of clusters fixed at two and starting with the cluster centers as the means of the points on each side of the dividing plane. The two cluster centers found in the procedure must, of course, be about 4σ apart in the pattern space if the two clusters so formed are to be considered readily separable in accordance with the previously established criteria.

This should break the large cluster into two smaller clusters

which may then be tested to see if they have the same shape as the clusters thought to represent single symbols. If the component clusters still exhibit a significant difference in the test, the procedure may be repeated on them until additional single-symbol clusters are isolated.

The isolation of single-symbol clusters is equivalent to proposing new symbols for the receiver to recognize, which symbols are suggested by the data itself. Thus, each time a new single-symbol cluster is isolated, the receiver augments the set $\{\bar{z}_j, q_j\}$, $j = 1,2,\ldots\ell$, by adding the symbol $\bar{z}_{\ell+1}$ and its relative frequency $q_{\ell+1}$. The mean of the cluster is used as the symbol estimate, and the sample relative frequency is taken as the estimate of $q_{\ell+1}$. Thereafter, the decision scheme is capable of recognizing received signals as representations of $\bar{z}_{\ell+1}$.

The estimation of the symbol relative frequencies, $\{q_j\}$, $j = 1,2,\ldots\ell$, is simplified because of the previous assumptions that symbols will not be recognized as distinct unless this can be done with a very low percentage of errors. The relative frequency of recognition of a symbol will generally not be equal to that symbol's relative frequency of transmission. If $\bar{R}$ is a vector of the relative frequencies of recognition of the various receiver symbols, and $\bar{T}$ is the vector of true probabilities of transmission of the transmitted symbols, the relationship between these vectors will be a transformation readily identified as the system transition matrix $\bar{P}$. The equation is $\bar{R} = \bar{P}\,\bar{T}$.

$$\begin{bmatrix} R_1 \\ R_2 \\ \vdots \\ R_\ell \end{bmatrix} = \begin{bmatrix} P_{11} & P_{12} & \cdots & P_{1K} \\ P_{21} & P_{22} & & \cdot \\ P_{31} & & & \cdot \\ \vdots & & & \cdot \\ P_{\ell 1} & \cdots & \cdots & P_{\ell K} \end{bmatrix} \begin{bmatrix} T_1 \\ T_2 \\ \vdots \\ T_K \end{bmatrix} \qquad \text{where } P_{ji} = \Pr(\bar{z}_j | \bar{x}_i)$$

Suppose that $\ell = k$, that is, the receiver has learned to identify all the transmitted symbols. Under the previously discussed assumptions, it is expected that P will be very much like the identity matrix except for permutations of rows and columns since P_{ji} will be almost unity for one pairing of the i's and j's and very small for all other combinations. Because of this, the $\bar{R}$ and $\bar{T}$ vectors are seen to contain very similar values except for their order in the vectors. That difference in order, or course, corresponds to the different ordering of the symbols at the receiver compared to the ordering at the transmitter. Thus, the relative frequency of recognition of a symbol at the receiver serves as a reasonably good estimate of the corresponding symbol's true probability of transmission.

The Utilization of Data in the Model

Only occasionally will one transmitted symbol be erroneously recognized as more than one symbol by the receiver as stated earlier: usually the converse will occur. Most of the problems encountered in recognizing all the individual transmitted symbols will be due to the inability of the system to separate in a satisfactory manner clusters which are combinations of single-symbol clusters.

The solution to such a problem is to develop imaginative decision schemes and/or schemes which utilize more of the available information. In the present situation, where the individual point clusters are distributed as $N(\bar{x}_i, \sigma^2 I_n)$, the maximum likelihood decision process implemented by using hyperplanes to partition the pattern space is known to be optimum so a search for more complex decision schemes using the same information would be pointless. The obvious thing to do in this case is to increase the number of symbol features utilized by the receiver and thereby take advantage of more information upon which to make the decisions.

The addition of more symbol features for use in the decision-making section of the receiver may be considered as an increase in the dimensionality of the pattern space. There are several reasons for such an interpretation. First, it is a natural extension of the concept of visualizing the received signals as points in the pattern space and the decision scheme as a set of planes which partition that space. Second, it is easy to see how augmenting the dimensionality of the pattern space permits the use of decision schemes yielding very much improved recognition of the transmitted symbols with only a simple extension of the decision-making process. Third, the received signals themselves may reveal how to perform the recognition process much more readily when the data is presented in the proper number of dimensions.

Figure 9 is an example of a situation where, using only features f_1 and f_2, the clusters of points generated by two transmitted symbols $\bar{x}_1$ and $\bar{x}_2$ overlap to such an extent that reliable recognition of the symbols is impossible. In two dimensions, the means of the clusters are only about 2σ apart so, if the symbols are equally likely, it is

impossible to differentiate between them at the receiver with less than about 16% errors. It is evident from the shape of the overall cluster that it is composed of at least two smaller clusters, but the system will not separate those clusters unless it can do so in a manner that satisfies its criterion of low recognition error rate.

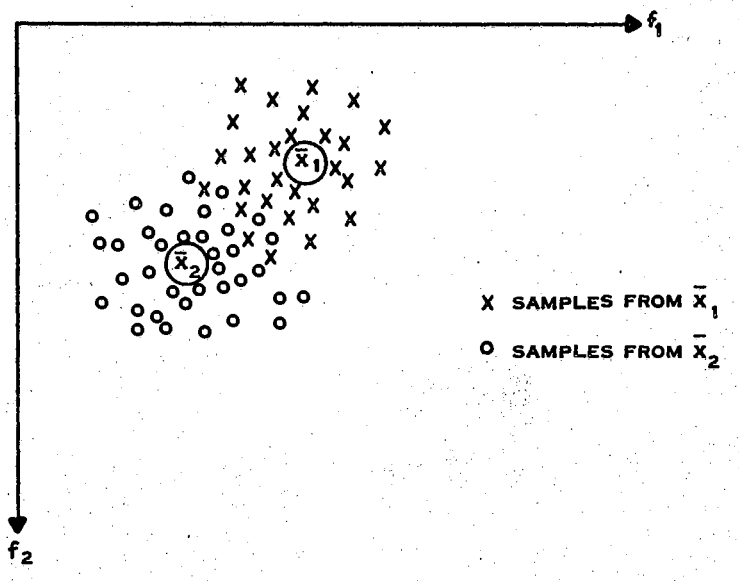

Figure 9. Samples in two Dimensions

In Figure 10 where the feature f_3 is considered, the situation changes considerably. The clusters may easily be separated in a three-dimensional decision scheme that is conceptually no more complicated than that used in two dimensions. The two clusters so found are far enough separated in the three-dimensional pattern space that the transmitted symbols generating the cluster may be easily recognized with low error probability. Thus, the receiver has learned to distinguish the

two transmitted symbols.

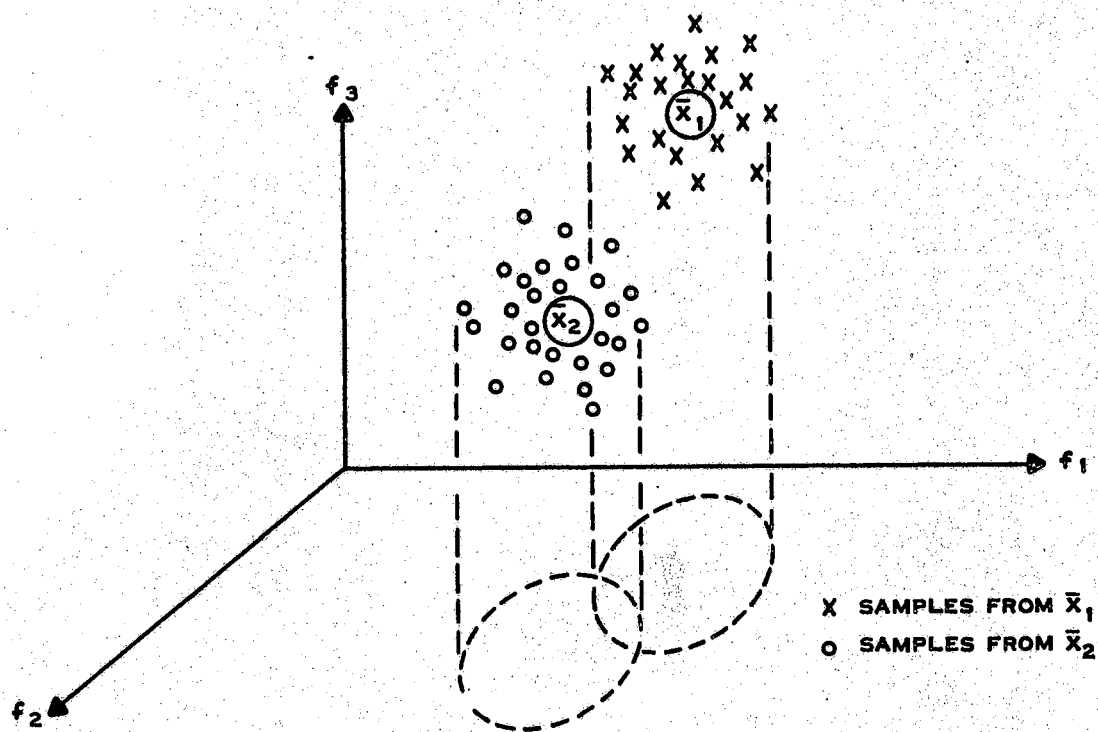

Figure 10. Samples in Three Dimensions

In retrospect, it is evident that the feature f_3 contains a relatively large amount of information about which of the two symbols, $\bar{x}_1$ or $\bar{x}_2$, was sent and this feature should surely be used in the recognition process. However, recall that at the outset of the learning process the receiver was not even aware of the existence of the two transmitted symbols, and so it could hardly be expected that the system

would know a priori which features are important for recognition of those symbols.

At this point, there remains one important aspect of the system operation which has been considered only briefly. That is the prospect that the receiver will attempt to recognize more different symbols than the transmitter is sending. In general, the difficulty here will be caused by the receiver's attempt to break a single-symbol cluster down into two or more component clusters. This can happen if the covariance matrix comparison test erroneously indicates a significant abnormality in the shape of a cluster. If the significance level of this test is made rather high, say 10%, in hope of detecting all discrepancies in cluster shape, it is expected that the test will indicate anomalous differences with that same frequency. That is, in 10% of the tests a significant difference will be indicated even when all clusters have been drawn from distributions with identical covariance matricies. Detecting such a shape discrepancy automatically starts the search for ways of dissecting the large cluster into its single-symbol components.

When a cluster is erroneously indicated to be composed of samples drawn from more than one transmitted symbol, the search for methods of isolating the single-symbol clusters is always due to fail since there is, in fact, no way of using linear discriminants to break a true single-symbol cluster into two or more smaller clusters to have the same size and shape as the original cluster except for noise. Even when all the other symbol features are examined in an attempt to find significant measurements upon which to differentiate between two symbols supposedly composing the cluster, no truly significant clusters will ever become apparent since all the sample features have, in fact, been drawn from

one single normal distribution.

Because of this, the receiver can be made very "inquisitive" in its search for symbols sent by the transmitter without much danger that it will consistently recognize more symbols than are actually present.

It can be seen that some minimum knowledge of the channel-induced noise is essential to the search for new symbols if the receiver is to be successful. First, the shape of the noise distribution should be known, at least insofar as the number of modes, so one can have some a priori knowledge of how the sample points should be grouped into clusters. Second, it is important to have some information about either the noise variance or the number of clusters to be found. Without any constraints offered by this a priori knowledge, it would be permissible to group each sample point into its own individual cluster, or else to group all the samples into one large cluster. Since the process of isolating previously unrecognized single-symbol clusters depends heavily on being able to find at least one single-symbol cluster during the initial clustering phase, some constraints of this type appear to be very useful. In the experimental situations to be considered in this study, the noise is always assumed to be distributed as $N(\emptyset,\sigma^2 I_n)$ with σ known so the necessary information is readily available. It is believed that the system could operate almost as effectively if the noise distribution were known only approximately.

CHAPTER III

ANALYSIS AND SIMULATION OF THE MODEL

The Computer Simulation

In order to test the model and to observe some of the learning phenomena alluded to in the introduction, the communication system model was simulated on a digital computer. The basic model of Figure 5 with the receiver of Figure 7 was simulated on an IBM System/360 using ALGOL as the programming language. Some features of the program are explained in the following section.

The only information supplied to the simulated transmitter is the set of transmitter symbols and their probabilities of transmission, $\{\bar{x}_i, p_i\}$. The channel is simulated by the addition of noise vectors, of known statistical form, to the transmitted symbols. The symbol selection process at the transmitter and the noise vector generating process in the channel are controlled by a pseudo-random number generator in the program. The receiver starts with only the knowledge of the channel noise statistics and the fundamental "idea" of how it is to operate.

The symbols selected by the transmitter are represented by vectors of real numbers, each transmitter symbol being represented by an unique n-dimensional vector where n is the maximum number of measurements, or features, characterizing the symbols. These n-dimensional transmitter symbol vectors $\bar{x}_i$ plus n-dimensional noise vectors generated by the simulated channel are presented to the receiver as signal

vectors $\bar{y}_r$ which the receiver attempts to recognize as representations of its receiver symbol vectors $\bar{z}_j$. All the vectors are n-dimensional except possibly the receiver symbol set $\{\bar{z}_j\}$ which is of dimensionality n_s, $n_s \leq n$, the number of signal features deemed necessary by the receiver for its recognition scheme. The receiver signal classification scheme starts as a partition of this n_s-dimensional signal space. This space may be augmented later as the receiver considers more signal information in the form of higher dimensional vectors. At the beginning of the process the receiver symbol set $\{\bar{z}_j\}$ is empty.

As pointed out previously, the receiver is able to recognize only the symbols of which it is "aware", that is the symbols represented by the set of vectors in the receiver symbol set $\{\bar{z}_j\}$. Therefore, when the first signal vector is presented to the receiver, it is not recognized as any existing receiver symbol since there are none. Instead, the signal is identified as representing a new symbol and an estimate of that symbol is placed in the receiver symbol set. The best estimate of that symbol is, of course, the signal that was presented to the receiver. Then the receiver symbol set consists of only one symbol representation vector $\bar{z}_1$, that being the first signal vector received.

When the second signal vector is presented to the receiver the situation is somewhat different. The second signal must be recognized by the receiver as a representation of the symbol $\bar{z}_1$ since that is the only symbol whose recognition is possible. On the other hand, there is the possibility that the transmitter selected a symbol different from the first symbol for transmission this time. To account for this possibility, the receiver must test the hypothesis that the second signal arose from the transmission of a symbol different from the first

transmitted symbol. That is, it tests the "differentness" of the second received signal from the symbol $\bar{z}_1$. If the test shows a significant difference, a difference not likely caused by the channel noise alone, the second received signal is identified as representing a new symbol and an estimate of that symbol is added to the receiver symbol set.

The receiver progresses in this manner, recognizing each received signal as one of the symbols $\{\bar{z}_j\}$ and then testing to see whether or not it is likely a new, previously unrecognized symbol by a measure of the difference between the actual received signal and the symbol as a representation of which it is recognized. This two-step process is the most efficient method of identifying new symbols since the recognition process picks out the receiver symbol most like the signal in question. If the signal is different enough from that receiver symbol to be identified as a new symbol, it is also different enough from all the other receiver symbols to lead to the same conclusion.

If it is decided that the signal represents a previously recognized symbol, the estimate of the recognized symbol is updated by averaging in the newly received signal, thereby improving the symbol estimate.

The symbol probabilities required by the recognition process are estimated by the observed relative frequencies of recognition of each symbol in the receiver symbol set. These estimates converge in probability to the true probabilities as the number of received signals grows.

It is expected that the new-symbol identifying test will generate anomalous new symbols with probability equal to the significance level of the test. For instance, transmitted symbol $\bar{x}_1$ which is properly recognized by the receiver as, say, $\bar{z}_5$ may upon one occurrence be

identified as a new symbol, say, $\bar{z}_{10}$. In order to assure that the receiver symbol set converges to the transmitter symbol set, it will eventually be necessary to identify $\bar{z}_{10}$ as a spurious representation of symbol $\bar{z}_5$ and to combine $\bar{z}_{10}$ with $\bar{z}_5$ using a weighted average of the two to form a new symbol $\bar{z}_5$. This is accomplished in the simulation by a statistical test referred to as the symbol combining test.

Each time a received signal is recognized as a representation of one of the previously existing receiver symbols, and used to update a symbol estimate, the symbol-combining test checks the "differentness" of the updated symbol estimate from all the other symbols. If two symbols are found to be very similar, it is assumed that one of them is a spurious representation of the other and the two are combined in a weighted average. Thus, each time a receiver symbol is changed by the addition of new data in its estimate, it is checked to be sure it is still different enough to be considered unique.

It was a preliminary assumption that the transmitter symbols would be different enough to be readily identifiable at the receiver when the proper recognition scheme was used. There are two reasons, then, that different transmitter symbols may be recognized as identical by the receiver. First, the symbol-combining test may erroneously indicate that two symbols are similar enough to be combined. This will happen with probability equal to the significance level of the test. Second, the receiver's recognition scheme may not utilize the proper features of the received signals to recognize as distinct all of the different symbols represented by those signals.

There is a third test applied to all the received data periodically to detect combinations of symbols caused by the occurrence of both the

above events. The dispersion of the signals recognized as a particular receiver symbol should be due only to the channel noise since the transmitted symbol should have been the same for each signal. If, in fact, the signals were generated by the transmission of more than one symbol, the dispersion of the signals will be greater than would be expected for the single symbol case. The third test, referred to as the cluster shape test, is designed to detect receiver symbol clusters whose dispersion is greater than would be expected if all the component signals represent only one transmitter symbol.

The determinant of the sample covariance matrix of the signals recognized as each receiver symbol is used as a measure of the dispersion of those signals. Presumably, the receiver symbol whose signals have the smallest dispersion represents a single transmitter symbol (provided a sufficient number of these signals have been received to justify considering them a cluster). Those receiver symbols whose signals have significantly larger dispersions are suspected of being the result of recognizing more than one transmitter symbol as one symbol at the receiver.

When such a receiver symbol is detected by the cluster shape test, an attempt is made to distinguish two or more transmitter symbols comprising the suspect receiver cluster. It may be possible to find a suitable partition of the signal space in n_s dimensions. A suitable partition is one which separates the received signal vectors according to the underlying transmitted symbols with low probability of error. If this is possible, the result is merely the generation of one or more new symbols to be added to the receiver symbol set. When a suitable partition cannot be found in this manner, it is necessary to seek to

utilize more information in the received signals by increasing the dimensionality of the space used in the receiver decision scheme. This is done by adding one dimension at a time until either n_s dimensions are reached or else a suitable partition is found. If a suitable partition is not found, it is concluded that the signals must have all been generated by the transmission of one symbol and that the cluster shape test results were in error. In this way the receiver is somewhat protected against excessive errors in the cluster shape test.

As the number of signals presented to the receiver grows, new symbols are discovered in the data and added to the receiver symbol set. In order to make the process efficient with respect to the number of signals required to form good estimates of the transmitter symbols, all of the received signals are periodically reviewed by the receiver as if each one had just been received for the first time. This is done by using the clustering algorithm discussed previously to group all the signals into compact clusters about each of the symbols in $\{\bar{z}_j\}$. In effect, the receiver starts the recognition process over at the beginning except that it has some a priori knowledge of the symbols and their probabilities to aid it this time. It is believed that this step is not necessary for convergence of the receiver symbol set but it increases the efficiency so that fewer signals must be processed by the computer program to obtain good estimates of the transmitted symbols. If a very large number of signals were processed, the effects on the symbol estimates of early misrecognition would become negligable in relation to the effects of correctly recognized data, assuming that the recognition scheme is corrected by the learning process. The fact that the recognition scheme is based on the symbol estimates resulting from

the recognition scheme means that misrecognized symbols tend to continue to be misrecognized until the learning process triggered by its dissatisfaction with its evaluation of its performance, rectifies the situation. This point is examined in somewhat more detail later.

Data Available From the Simulation

Since it is the dynamic behavior of the learning receiver as well as its final amount of learning about the transmitter that is of interest in this investigation, the computer simulation prints out data indicating the status of the adaptive parts of the receiver frequently during the learning process. Messages are printed whenever a new symbol is generated, whenever two symbols are combined into one, and whenever the cluster shape test finds a symbol suspected of being a combination of two or more transmitter symbols. At regular intervals the program prints out a complete status report including the set of receiver symbols and their probabilities used by the recognition process and the covariance matrix of all clusters recognized as each symbol. Perhaps the most easily assimilated and therefore most meaningful indication of the system's behavior is the average information flow through the communication system model as a function of the number of symbols processed by the receiver. This information is printed in the form of a graph referred to as the learning curve. Such a curve is shown in Figure 11.

Most of the simulation results reported herein were obtained using a set of six transmitter symbols, each symbol being described by a vector of three real numbers. All symbols have the same number in the third element of their vector descriptions so actually only the first two elements are of any significant use in the recognition process.

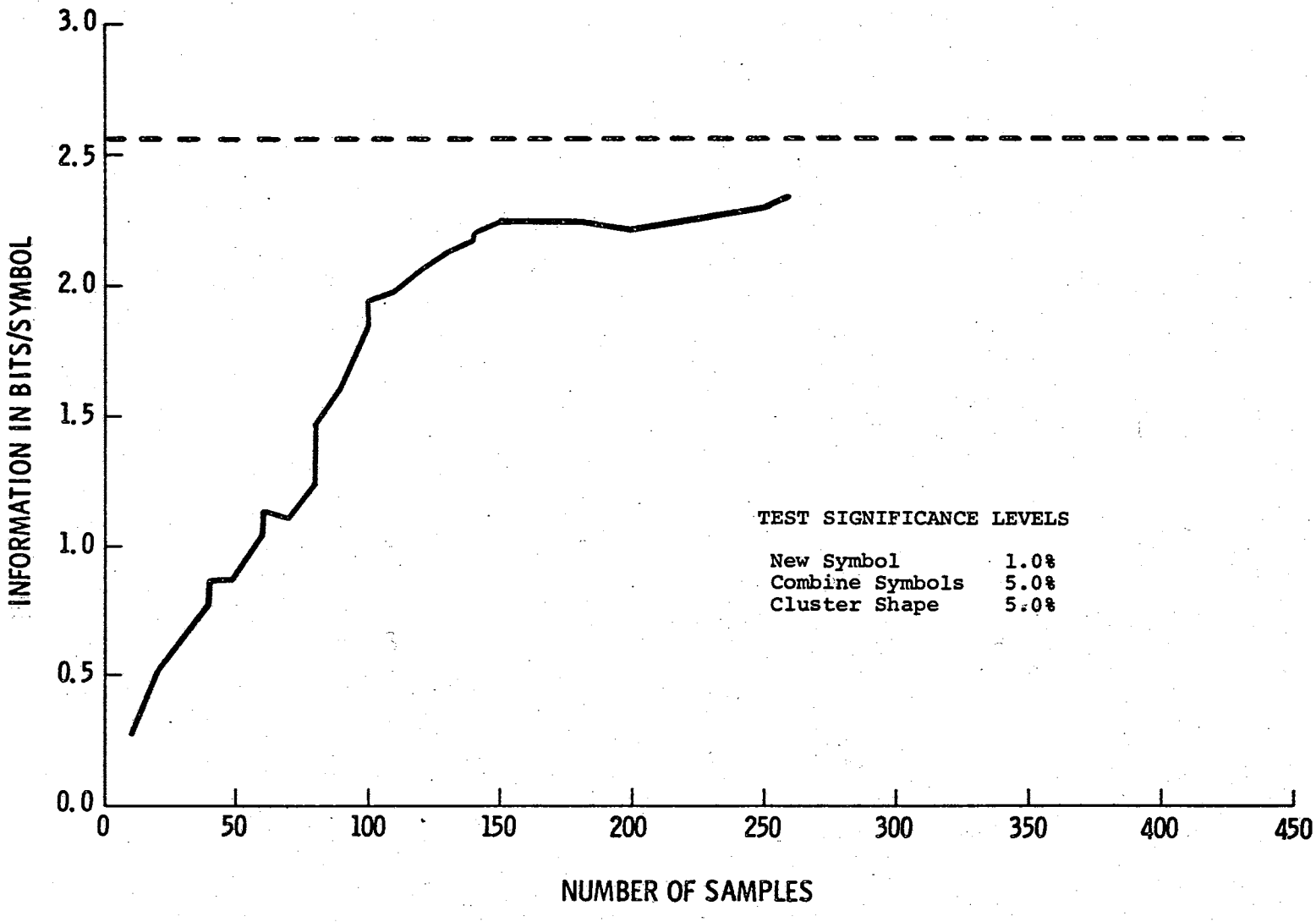

Figure 11. Typical Learning Curve

Figure 12 shows graphically the first two elements of each transmitter symbol along with the probabilities of transmission of each. As can be seen, the most similar two symbols are $\bar{x}_3$ and $\bar{x}_4$. If the channel noise has standard deviation of unity, these two symbols are about 5.4σ apart so they should be readily distinguishable by the proper recognition scheme in two dimensions at the receiver.

The receiver always starts by considering only the f_1 element of the signal vectors, in which signal space only symbol $\bar{x}_6$ is uniquely identifiable. In order to recognize the symbols $\bar{x}_1$ through $\bar{x}_5$ in a manner satisfying the criterion for low-error recognizability, it is necessary for the receiver to search for more information in the second and third elements of the signal vector.

The Measure of Learning by the Receiver

Measuring the state of knowledge of a learning system or the amount of learning which it has accomplished is a broad and loosely defined problem and, consequently, one which is subject to a great deal of individual interpretation. Since the purpose of learning in a goal-directed system, the type just described, is to permit that system to accomplish a predetermined task, it seems natural to measure the amount of learning by how well the system is able to perform the task. In the case of the chess and checker playing systems mentioned earlier where the goal is to win the game, measuring the ability of the system to perform its assigned task can be a formidable undertaking by itself. In general, the problem of measuring learning is neither well defined to begin with nor, once defined, easy to deal with because of the complexity of the systems to which the problem applies.

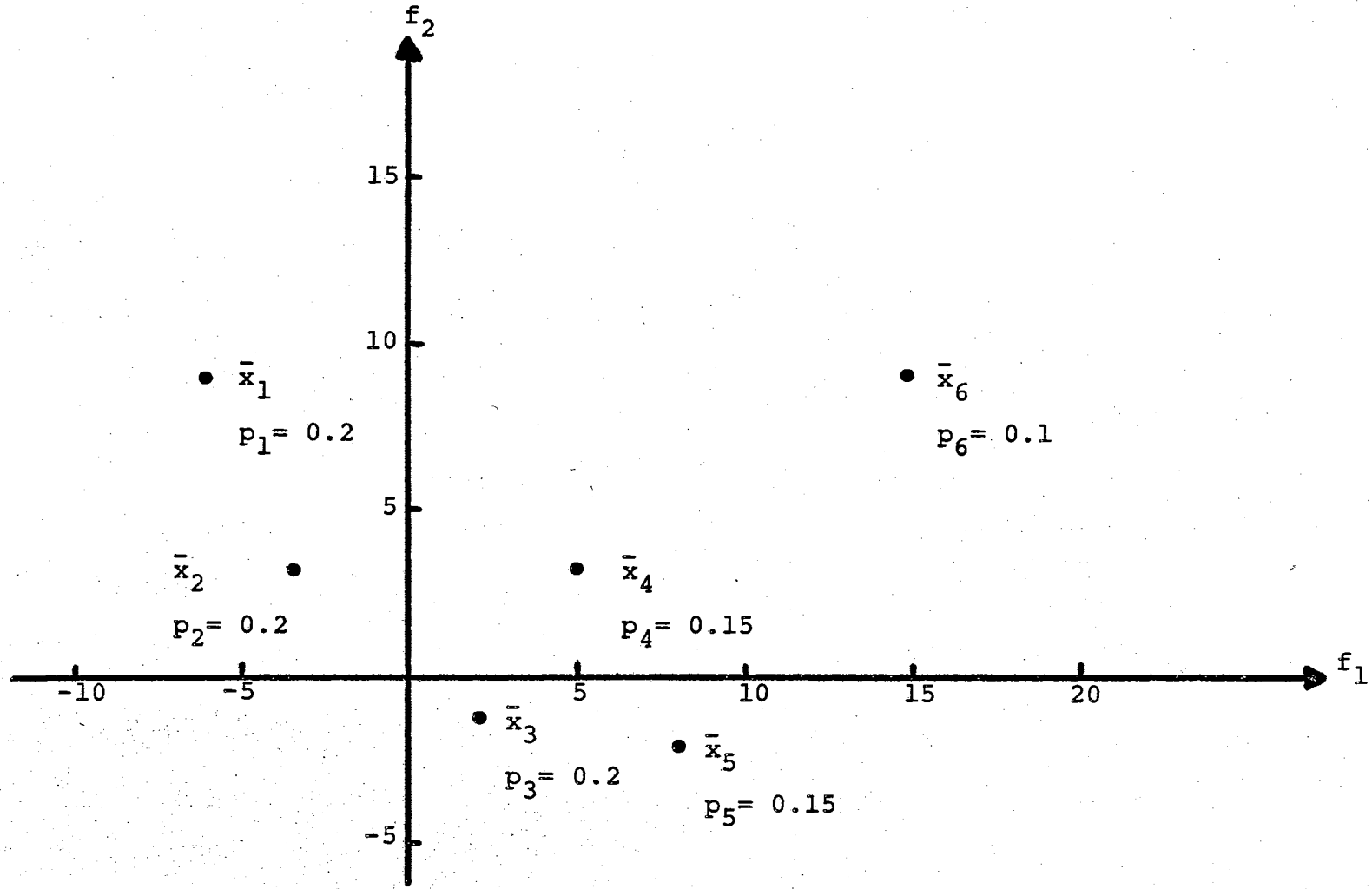

Figure 12. Transmitter Symbol Set

In the communication system studied here, a relatively simple learning system, the goal of the receiver is to recognize each of the transmitted symbols distinctly and with low error probability. This will be accomplished as completely as possible when the receiver symbol set $\{\bar{z}_j, q_j\}$, $j = 1,2,\ldots\ell$, has a one-to-one correspondence with the transmitter symbol set $\{\bar{x}_i, p_i\}$, $i = 1,2,\ldots k$, and $p_i = q_j$ for the pairings achieved. At the other end of the scale, we take as the worst performance of the assigned task that which occurs when the receiver symbol set is $\{\bar{z}_1, 1\}$; that is, when every received signal will be recognized as the same symbol with unity probability.

In light of this, it would be convenient to measure learning in the communication system by a comparison of the receiver and transmitter symbol sets. Unfortunately, there are several inherent difficulties in this approach. For one thing, the decision scheme implemented at the receiver does not depend, in the general case at least, on a unique set of $\bar{z}$'s and q's; identical decision schemes may be generated by more than one different receiver symbol set. Second, the receiver decision scheme is to be based on only the features of the received signals which are important to the recognition of the different symbols and not on all available information. Thus, it is not necessary to have $\bar{z}_j = \bar{x}_i$ for some particular i and j; it is sufficient merely to have the proper features of the $\bar{z}_j$ equal to the corresponding features of the $\bar{x}_i$. For this reason, a knowledge of which features are important to the recognition process is necessary when evaluating the difference between transmitter and receiver symbol sets, and a meaningful measure of the importance of individual features may be difficult to find. Third, when comparing the sets $\{\bar{z}_j, q_j\}$ and $\{\bar{x}_i, p_i\}$ one must deal with some relation

between the i's and j's; that is, it must be known which $\bar{z}_j$ is the receiver's concept of the transmitted symbol $\bar{x}_i$ for every i and j. In the learning system it is quite possible that the relation between the i's and j's may change as learning progresses. For instance, it may be that $\bar{x}_2$ is recognized as $\bar{z}_4$ by the learning receiver in the early stages of learning, but later the receiver's decision scheme may be changed in the learning process so that $\bar{x}_2$ is recognized as, say, $\bar{z}_6$. These changes in the relationship between the transmitter and receiver symbol sets add a great deal of complexity to any such comparisons.

Because of these difficulties, it seems best to resort to a more classical measure of system performance, information flow, as an indication of learning. The measure of information flow in a communication system as an indication of system performance as proposed by Shannon (12) has usually been applied to non-adaptive systems and some extension of the ideas will be necessary if the application is to be meaningful in the context of a learning system.

In the Shannon model, transferred information is equal to the amount of resolution of uncertainty at the receiver about the transmitted symbols; that is, information transferred by the system is equal on the average to the difference in uncertainty about the transmitted symbols before and after reception of the transmitted symbols. The amount of information conveyed to the receiver by the reception of a particular symbol is a function of the unexpectedness of that symbol. The more unexpected a symbol is, i.e., the lower the probability of its reception, the greater the information conveyed by its recognition.

It is assumed in the Shannon model that the receiver possesses complete knowledge of the probabilities of transmission of each symbol

and the channel characteristics so the receiver's expected reception probability for each symbol is equal to the true probability. This corresponds to the situation in the learning receiver when $\{\bar{z}_i, q_i\} = \{\bar{x}_i, p_i\}$, $i = 1,2,\ldots k$. Statistical knowledge of the channel then permits the calculation of the transition probabilities, $P(\bar{z}_j|\bar{x}_i)$ $i,j = 1,2,\ldots k$, and the average information transferred per symbol may be calculated as:

$$I = H(\bar{X}) - H(\bar{X}|\bar{Z})$$

where

$$H(\bar{X}) = \sum_{i=1}^{k} p_i \log_2 \frac{1}{p_i}$$

$$H(\bar{X}|\bar{Z}) = \sum_{i=1}^{k} \sum_{j=1}^{k} P(\bar{z}_j|\bar{x}_i) p_i \log_2 \frac{1}{P(\bar{x}_i|\bar{z}_j)}$$

$$P(x_i|z_j) = \frac{P(\bar{z}_j|\bar{x}_i)p_i}{\sum_{m=1}^{k} P(\bar{z}_j|\bar{x}_m)p_m} \quad .$$

$H(\bar{X})$ is the entropy of the transmitted signal or average measure of uncertainty of a transmitted symbol, and $H(\bar{X}|\bar{Z})$ is the same measure of uncertainty when the symbol recognized by the receiver is known. The difference of these two functions gives a measure of the average information transferred per symbol.

$$I = \sum_{i=1}^{k} \sum_{j=1}^{k} P(\bar{z}_j|\bar{x}_i) p_i \log_2 \frac{P(\bar{x}_i|\bar{z}_j)}{p_i} \quad .$$

An alternate formulation for the average information transferred

per symbol is:

$$I = H(\bar{Z}) - H(\bar{Z}|\bar{X})$$

$$= \sum_{i=1}^{k} \sum_{j=1}^{k} P(\bar{z}_j|\bar{x}_i)p_i \log_2 \frac{P(\bar{z}_j|\bar{x}_i)}{P(\bar{z}_j)}$$

where $H(\bar{Z})$ is the entropy of the random variable $\bar{Z}$ and $H(\bar{Z}|\bar{X})$ is the entropy of $\bar{Z}$ when $\bar{X}$ is known. The two formulations are, of course, equivalent but the first emphasizes the uncertainty about the transmitted signal and the second is written in terms of the uncertainty in the received signal.

For the general case of the learning receiver, however, the analysis of information flow is not quite so straightforward because the amount of information conveyed to the receiver through its recognition of a particular symbol is subject to some degree of personal interpretation. In the first place, the subjective symbol probabilities used by the receiver may be considerably at variance with the relative frequencies of the symbols. In addition, the symbols expected by the receiver may be quite different from those sent by the transmitter. Because of these factors, the various entropies by which information flow is measured can be formulated in several different ways according to the particular interpretation used, each yielding different values for the information flow.

In the Shannon model there is a symmetry of information flow between the transmitter and receiver through the channel. If the functions of the transmitter and receiver are reversed, the information flow backward through the channel will be equal to that found for the forward case. This is evidently not directly applicable to the situation

of a learning receiver such as investigated here and, in fact, there is good reason to think that the information flow should be unsymmetrical in such a system. Part of the information transmitted to a learning receiver is lost in the channel as expected but part is also lost in the receiver itself because of its state of knowledge. Some of the received information should be used by the receiver to improve its state of knowledge.

The method of defining information flow used in this paper seems to serve the purpose for which it is utilized but no claim of uniqueness is made and there is no assurance that more meaningful methods do not exist.

The system is analyzed from the point of view of an outside observer as a communication system in which the transmitted symbols are selected from $\{\bar{x}_i\}$, each with probability p_i, and the received signals are recognized as members of $\{\bar{z}_j\}$ according to the decision scheme then in use at the receiver. The channel characteristics are assumed known.

With the knowledge of the $\bar{x}$'s, p's, $\bar{z}$'s, the channel characteristics, and the receiver decision scheme, one may calculate the transition probabilities $P(\bar{z}_j|\bar{x}_i)$ $i = 1,2,\ldots k$, $j = 1,2,\ldots \ell$. The average transferred information per symbol is then:

$$I = H(\bar{X}) - H(\bar{X}|\bar{Z})$$

where

$$H(\bar{X}) = \sum_{i=1}^{k} p_i \log_2 \frac{1}{p_i}$$

$$H(\bar{X}|\bar{Z}) = \sum_{i=1}^{k} \sum_{j=1}^{\ell} P(\bar{x}_i,\bar{z}_j) \log_2 \frac{1}{P(\bar{x}_i|\bar{z}_j)}$$

$$P(\bar{x}_i, \bar{z}_j) = P(\bar{z}_j | \bar{x}_i) p_i$$

$$P(\bar{x}_i | \bar{z}_j) = \frac{P(\bar{z}_j | \bar{x}_i) p_i}{\sum_{m=1}^{k} P(\bar{z}_j | \bar{x}_m) p_m} \quad .$$

So

$$I = \sum_{i=1}^{k} \sum_{j=1}^{\ell} P(\bar{z}_j | \bar{x}_i) p_i \log_2 \frac{P(\bar{z}_j | \bar{x}_i)}{\sum_{m=1}^{k} P(\bar{z}_j | \bar{x}_m) p_m} \quad .$$

The subjective probabilities $\{q_j\}$, the receiver's estimate of the received symbol probabilities, do not enter directly into the information flow equation except through the transition probabilities $P(\bar{z}_j | \bar{x}_i)$. These probabilities are determined in part by the receiver decision scheme which, in turn, is directly affected by the $\{q_j\}$.

It is the transition probabilities which change as the system learns, and so all the information about changes in the state of learning by the receiver is contained in them. As the receiver learns to recognize each of the transmitted symbols it is expected that $P(\bar{z}_j | \bar{x}_i)$ will approach 1 for those i-j pairs which define the relation between the transmitted and received symbols; $P(\bar{z}_j | \bar{x}_i)$ will be very small for all other i-j pairs. Knowledge of this i-j pairing is not necessary for calculation of I, nor is there any need for specific information about the symbol features used by the receiver. This information is manifest in the $P(\bar{z}_j | \bar{x}_i)$'s.

The calculation of those probabilities is reasonably straightforward in theory. The pattern space is partitioned into ℓ decision regions S_j, $j = 1,2,\ldots\ell$, where

$$S_j = \{\bar{y} | f_{\bar{Y}|\bar{X}}(\bar{y}|\bar{z}_j)q_j > f_{\bar{Y}|\bar{X}}(\bar{y}|\bar{z}_i)q_i, \qquad i = 1,2,\ldots k\} \quad .$$

Then

$$P(\bar{z}_j|\bar{x}_i) = \int_{S_j} f_{\bar{Y}|\bar{X}}(\bar{y}|\bar{x}_i)d\bar{y} \quad .$$

In practice, the evaluation of this integral presents a problem unless the vector space is of only one or two dimensions. Since it is desired here to be able to deal with much higher numbers of dimensions, an approximate method of evaluating the probabilities will be employed. The probabilities will be approximated by a bell-shaped curve which is algebraically manageable. Let

$$P(\bar{z}_j|\bar{x}_i) \doteq \frac{\dfrac{q_j}{5\, d^2(\bar{z}_j,\bar{x}_i) + 0.05}}{\sum_{m=1}^{\ell} \dfrac{q_m}{5\, d^2(\bar{z}_m,\bar{x}_i) + 0.05}}$$

where $d(\bar{z}_j,\bar{x}_i)$ is the Euclidean distance between the points $\bar{z}_j$ and $\bar{x}_i$ in the pattern space. This is, to be sure, a rough approximation but for the cases to be considered in this investigation, where $\bar{x}_i$ is close to one of the $\bar{z}$'s and relatively far away from all the other $\bar{z}$'s, it will serve the purpose for calculating an approximation to the information flow in the system.

For a fixed transmitter symbol set, the average information transferred per symbol is limited by the channel noise. This limit will be reached, for the most general case, only when the receiver has completely learned the transmitter symbols and their probabilities. When $\{\bar{z}_j,q_j\} = \{\bar{x}_i,p_i\}$ the average information transferred per symbol is:

$$I = I_{max} = \sum_{i=1}^{k} \sum_{j=1}^{k} P(\bar{x}_j | \bar{x}_i) p_i \log_2 \frac{P(\bar{x}_j | \bar{x}_i)}{\sum_{m=1}^{k} P(\bar{x}_j | \bar{x}_m) p_m} .$$

Under all other conditions, the information flow will not exceed this limit[1].

The average transferred information per symbol as a function of the number of symbols processed through the system gives a reasonably useful measure of how rapidly learning takes place in the learning receiver. This measure is plotted in the form of a graph like Figure 11 from which some inferences about the system's operation can be made.

The Learning Curves

As indicated, the behavior of the simulated system may be observed in a number of ways. The learning curve presents a view of the overall level of receiver performance as a function of the number of signals processed from which it is relatively easy to observe the effects of changes in system parameters. The curve of Figure 11 shows how the information flow, the observable measure of learning, increases rapidly at first as the receiver forms a rough estimate of the transmitter symbols to use in the recognition scheme. The final stages of learning take place more slowly as the estimates are refined by the reception of a growing body of data. Generally it will be convenient in this study to consider the learning to take place in two parts; first, the initial generation of the transmitter symbol estimates and, second, the final

---

[1] This assertion is easy to prove if it is also assumed that the $P(\bar{x}_j | \bar{x}_i)$ for $j \neq i$ are all equal (13). It has not been proved in general.

refinement of those estimates. In the learning curve, the change from the first to the second stage is denoted by the knee of the curve.

The learning processes investigated here are observed to occur primarily in the early part of the receiver's experience. During the first several hundred samples the receiver learns rapidly but somewhat inconsistently. The slope of the curve and its smoothness are generally the external features of interest here. Actually, of course, it is the way in which the receiver learns the transmitted symbols and their probabilities which determines the shape of the curve here and it is the effect of this underlying mechanism of learning that is to be observed in the learning curve.

The refinement of the symbol estimates which takes place during the second stage of receiver learning depends heavily upon the initial estimates found during the early learning part. If the initial estimates are reasonably good, it is expected that the information flow curve will smoothly and steadily approach the limit imposed by the transmitter and channel as the symbol estimates are gradually refined. When the initial symbol estimates are poor, and particularly when the number of transmitted symbols is chosen incorrectly by the receiver, the learning curve exhibits much less consistent behavior. Generally, it will be convenient to consider the first stage of learning to end only when the symbol estimates are sufficiently accurate to enable the receiver to recognize the transmitted symbols with a reasonably low probability of error. For the most part, large changes in the receiver symbol set and the generation of new receiver symbols will occur only during the first learning stage.

As discussed earlier, there are three variable system parameters

which may be externally controlled to adjust the receiver's performance. These are:

(1) the significance level of the test to decide when a single received signal represents a previously unrecognized symbol;

(2) the significance level of the test to decide when two symbols recognized as distinct by the receiver should instead be recognized as only one symbol;

(3) the significance level of the test to decide when a group of signals supposed by the receiver to represent only one symbol should actually be recognized as two or more distinct symbols.

From a broad point of view, one would expect that the first parameter should have the most noticeable affect on system performance during the very early learning stages when the receiver has not been "exposed" to all of the transmitted symbols. When this parameter is near its optimum value, the receiver should recognize new transmitted symbols with little delay. When this threshold is too high or too low, the receiver might erroneously indicate new symbols where there are none, or else fail to recognize new symbols as new when they are received. Similarly, the second parameter controls the ability of the receiver to discover that it has incorrectly distinguished two or more different symbols where there is, in fact, only one. If this parameter is set too high or too low, the receiver may erroneously recognize many different received symbols as identical or else fail to identify situations in which several apparently different symbols should be recognized as identical.

The third parameter controls the receiver's ability to look at all the signals received up to that time and to discover differences in the way the noise and the recognition scheme has affected each symbol recognized in the signals. Since the noise is presumed to affect each signal in a statistically similar manner, significant differences among the recognized symbols are assumed to be caused by improper recognition processes. This causes the receiver to re-examine the parts of the recognition scheme which led to the identification of the suspect symbols in hope of discovering new recognition criteria. This facet of the receiver's operation should affect the learning curve after a sufficient number of signals have been received to form a good estimate of the effects of noise and the recognition process on each symbol. Further, it is necessary to be able to separate the effects of the recognition process from those of the noise on the individual symbols. Only when a sufficient number of signals have been received to permit this can the receiver alter its recognition scheme in a meaningful way, that is, in a non-random manner.

Generally, in the range of reasonably "good" settings of the adjustable parameters it should be expected that these parameters have their greatest effect on system operation during the early part of learning. After the receiver once forms a good estimate of the transmitted symbols it would take a major disturbance to cause a significant change in the recognition process. It is in the formation of the initial estimate of the transmitted symbols and the subsequent widely varying attempts at improvement that the effects of the parameters should be most noticeable. On the other hand, if the parameters are set so that the initial receiver symbol estimates are very much in error, the gradual refinement

process of the second learning stage will be ineffective. In this way the parameters can be said to affect the second learning stage although it will usually be more convenient to think of this as a secondary reaction.

CHAPTER IV

RESULTS OF THE SIMULATION

By far the bulk of the results obtained from the simulation of the learning communication system pertains to the set of transmitter symbols described graphically in Figure 12. Because of the large amount of computer time necessary to simulate a single learning experience, about 30 minutes to simulate the receiver's processing of 350 signals, it was difficult to obtain a large enough body of data to permit a thorough analysis of the many different aspects of the system's operation. In concentrating on a single set of input data, it was hoped that enough about the system's operation for that one set of conditions could be learned and understood in order to indicate more promising areas for investigation.

The general procedure was to start the receiver from a state of complete ignorance about the transmitted symbols and to observe the progress of learning. The procedure was repeated for as many different sets of the three test parameters as possible. It was hoped that the data thus obtained could be used to derive some understanding of the system's dependence on these parameters and how they work together and individually to affect the learning phenomenon.

About half of the 30-odd computer runs were made with identical sequences of data presented to the receiver. Most of the remainder of the test data was the result of using the same transmitter symbol set

but allowing the transmitter to generate different random sequences of those symbols. Some brief experimentation was done with more complex transmitter symbol sets utilizing various geometrical configurations and higher numbers of dimensions.

As indicated previously, most of the results of the simulation are presented in the form of learning curves detailed with notes about the underlying system operation at points of interest. Just exactly how much can be inferred about the system's operation, not to mention the fundamental learning process, by examination of these curves is somewhat a matter for subjective interpretation. Nevertheless, this appeared to be the most meaningful data obtainable under the general limitations inherent in this study. Generally, the slope of the learning curve and its smoothness were the features deemed most important in these observations.

Perhaps not too surprising, the performance of the learning system was not greatly affected by the parameter values as long as they were within a rather broad neighborhood of "good" values. It was soon discovered that the system's performance on a given run was much more dependent upon the actual sequence of symbols than upon the parameter values. The generation of this sequence was controlled by a pseudo-random number generator within the program so the actual sequence was not directly under external control. However the starting point for the sequence was used to control whether the same or different sequences of symbols were to be used on different runs. In this way it was possible to separate to some degree the effects of changes in the receiver parameter values from changes in the sequence of data presented to the receiver.

In regard to the results of the simulation and what is to be inferred from those results, it is certainly worth making note of the fact that the simulated system works as expected. In all cases within a very broad range of conditions, the simulated receiver generated a symbol set and associated probabilities that formed a good approximation to those at the transmitter. The learning curve always tended to increase until it asmytotically approached the limit set by the transmitter and channel. Evidently the fundamental structure of the system and its simulation is sound.

Detailed Example of Learning Process

Figure 13 is a plot of the learning curve for a perhaps not too typical run but one which is good for showing in some detail the major changes and adjustments made in the receiver symbol set during the learning period. The transmitter symbols used are those shown in Figure 12. During the reception of the first 20 signals, the receiver was able to detect five of the six different transmitted symbols at the 0.1% significance level. Inspection of the printed output of the program reveals that at this point the receiver was recognizing symbols $\bar{x}_3$ and $\bar{x}_4$ as identical. The printed output also shows that symbol $\bar{x}_3$ had been sent only twice thus far by the transmitter so it is understandable that the new symbol had not been "noticed" by the receiver. Interestingly enough, this seemingly insignificant error caused the receiver to get off to a poor start from which it did not fully recover until considerably more data had been received. After 40 symbols had been received and classified, the clustering algorithm was used to regroup the signals into five compact clusters. This redefined the receiver symbols

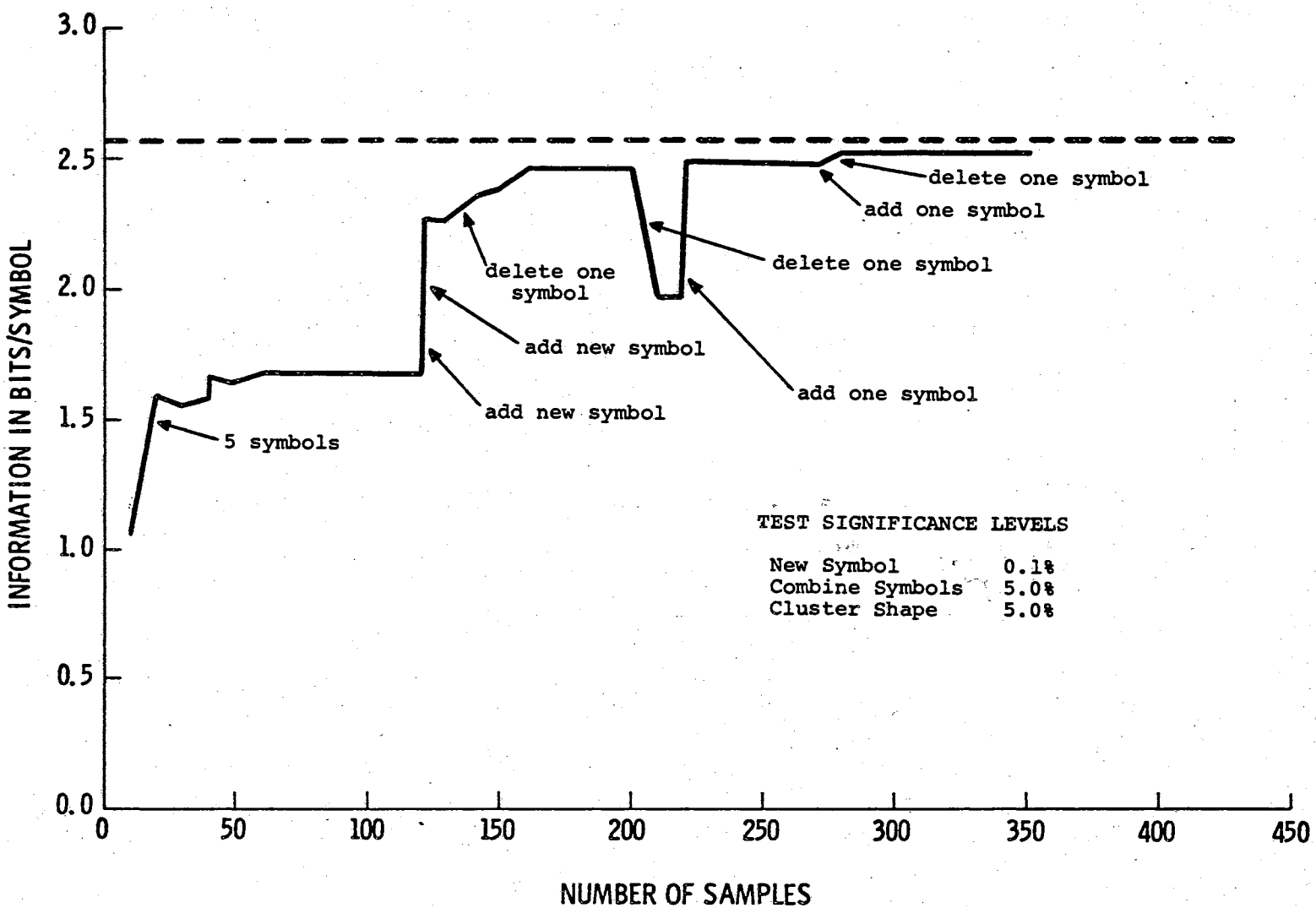

Figure 13. Example of Learning Process in the Simulation

slightly and resulted in a slight increase in average information flow. After 120 signals had been received, the cluster shape test detected a significant difference in the shapes of two of the clusters and caused both clusters to be subdivided, resulting in the recognition of two additional symbols by the receiver. The receiver was at this point recognizing seven distinct symbols but the transmitter was sending only six. Almost immediately, the receiver decided that the extra symbol was identical to one of the others and combined the two. The clustering algorithm, after 140 signals, again improved the average information flow slightly by refining the symbol estimate. At this point the receiver was recognizing the correct number of symbols and inspection of the printed output shows that only six of the 140 signals had been recognized incorrectly. The gradual improvement of the symbol estimates then resulted in a continually increasing information flow until, with about 200 signals processed, the receiver suddenly combined two of the symbols and then added a new symbol to its set of estimates. The clustering algorithm, at 220 signals, regrouped the signals into the most compact clusters starting with the existing symbol estimates for the cluster centers. This whole process resulted in only a very slight improvement in the symbol estimates since they were quite good to begin with, but the subsequent reclassification of all the symbols shows that only 4 of the 220 signals were then recognized incorrectly. Evidently the slight improvement was enough to help in the classification of the most questionable signals. After about 260 signals, the receiver erroneously recognized a new symbol and then combined one of the others with it. This did not have any noticeable effect on the average information per symbol because there was always at least six receiver symbols.

Before, when the receiver had combined two symbols and then proposed one new symbol, the information flow graph showed a large drop in average information flow during the time when there were too few receiver symbols. Generally, the addition of only one or two extra symbols to the receiver symbol set should not have a great effect on information flow because the estimated probability of reception of the extraneous symbols is ordinarily quite low. This time when the receiver settled down to receiving the correct number of symbols the printed output from the program shows that all the transmitted signals were recognized correctly by the receiver. From this point onward, the receiver recognized all of the transmitted symbols with no errors and there occurred no further disturbance of the recognition scheme.[1]

Although it is not at all evident from the learning curve itself and no notation is made on the graph, all the new symbols proposed by the receiver use two dimensions in the recognition scheme. At the beginning of the simulation process, only the f_1 variable of the signals were considered as a characterization of the transmitted symbol. As noted previously, this variable does not contain sufficient information for reliable recognition of the different symbols in this symbol set. The f_2 variable is considered by the receiver when it discovers that the classification scheme is unsatisfactory in one dimension but much improved in two dimensions. The final result in this example is that two dimensions are used to classify all the signals except those recognized as representations of symbol $\bar{x}_6$. Inspection of the transmitted symbol vectors shows that this is as it should be; symbol $\bar{x}_6$ is

---

[1]Extraneous symbols can be expected to be postulated at a rate about equal to the significance level of the first test.

separated by seven units from its closest neighbor in the f_2 direction and only 4.66 units separation is necessary to meet the receiver's criterion for distinguishability.

The receiver does not make use of the f_3 variable in any part of the recognition process. The printed output of the program reveals that the third dimension is inspected several times in a search for new symbols among clusters whose shapes are found questionable by the third test but this variable is each time rejected as a discriminant. Indeed, the f_3 element of the signal vectors varies only with noise and so cannot aid in the recognition process.

Effects of Parameter Changes in the Receiver

In order to obtain some meaningful data for the purpose of investigating the system's reactions to changes in the three variable parameters it was decided to make a number of runs using identical sequences of signals and changing only the parameters. It was reasoned that the differences in system operation from one run to the next could be observed and some conclusions drawn about the dependence of the learning process upon these parameters.

The limited amount of available data prevents making any but the most general types of statements in this regard. Perhaps the most general statement that can be made, and quite possibly the most meaningful statement also, is that the important part of the receiver's operation was essentially unaffected by rather large variations is the parameters. As long as the parameters were within a neighborhood of their "good" values, the simulated receiver would learn the transmitted symbol set with reasonable efficiency.

This is not to say that the learning system showed no reaction to changes in the parameters. To the contrary, even a slight shift in one parameter might have a very great affect on certain characteristics of the learning curve. The way new symbols are postulated and accepted or rejected controls the shape of the learning curve, and the precise manner in which the new symbols are postulated is controlled by the test parameters. Even so, the overall learning ability of the receiver was unaffected by these changes.

Figure 14 shows the learning curves of ten runs using identical data with various values for the three parameters. The identical sequences of data were generated by using the same starting point for the pseudo-random number generator for each run of the program. The transmitter symbol set was that shown in Figure 12. The significance levels of the three tests were varied over ranges expected to make evident any dependency of the learning process on these parameters. For this particular sequence of data the cluster shape test significance level seems to be rather uncritical in the whole range tried from 1% to 10%. This is an intuitively pleasing result since it is believed that the search for new symbols triggered by this test will only rarely result in the generation of spurious receiver symbols. The effects of changes in this parameter generally do not show in the learning curves since the receiver symbol set is not influenced by unproductive searches for new symbols. As long as this test is sensitive enough to detect abnormalities which may lead to the discovery of a new symbol, the learning process should be practically independent of this parameter. On the other hand, the number of unproductive searches, or "false alarms", should increase directly with this significance level. The

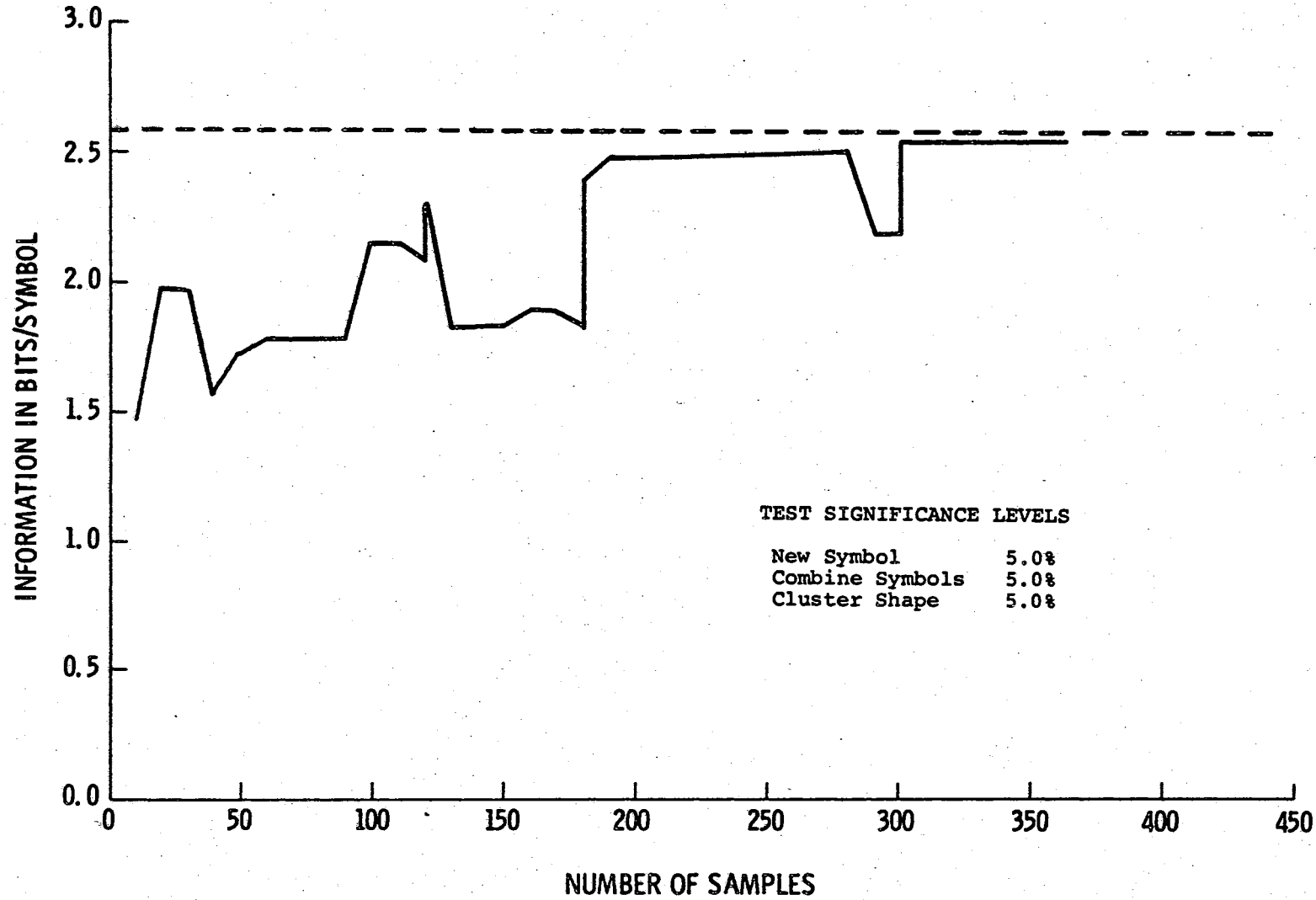

Figure 14. Learning Curves Showing Effects of Changes in Simulated Receiver Parameters

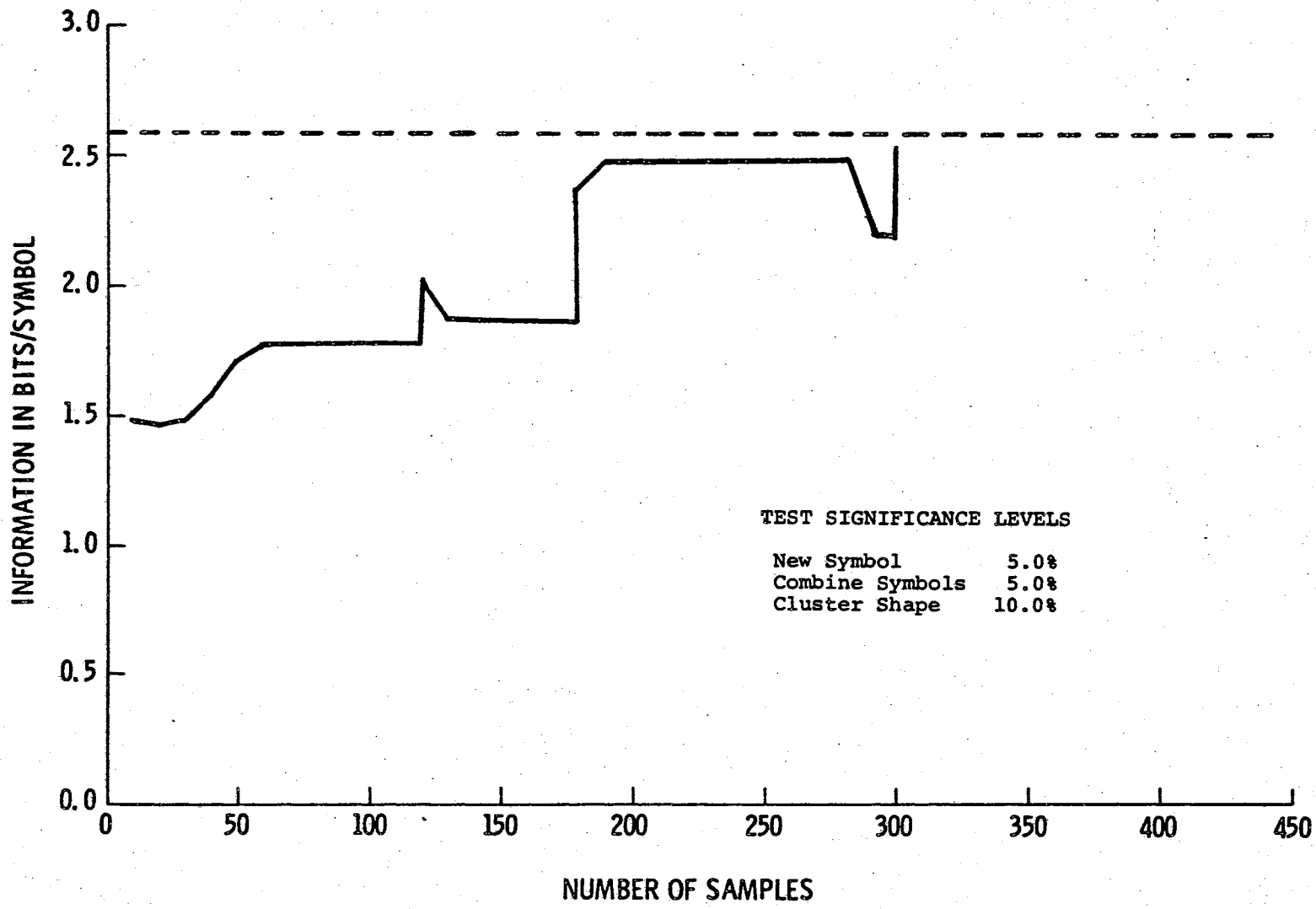

Figure 14 (Continued)

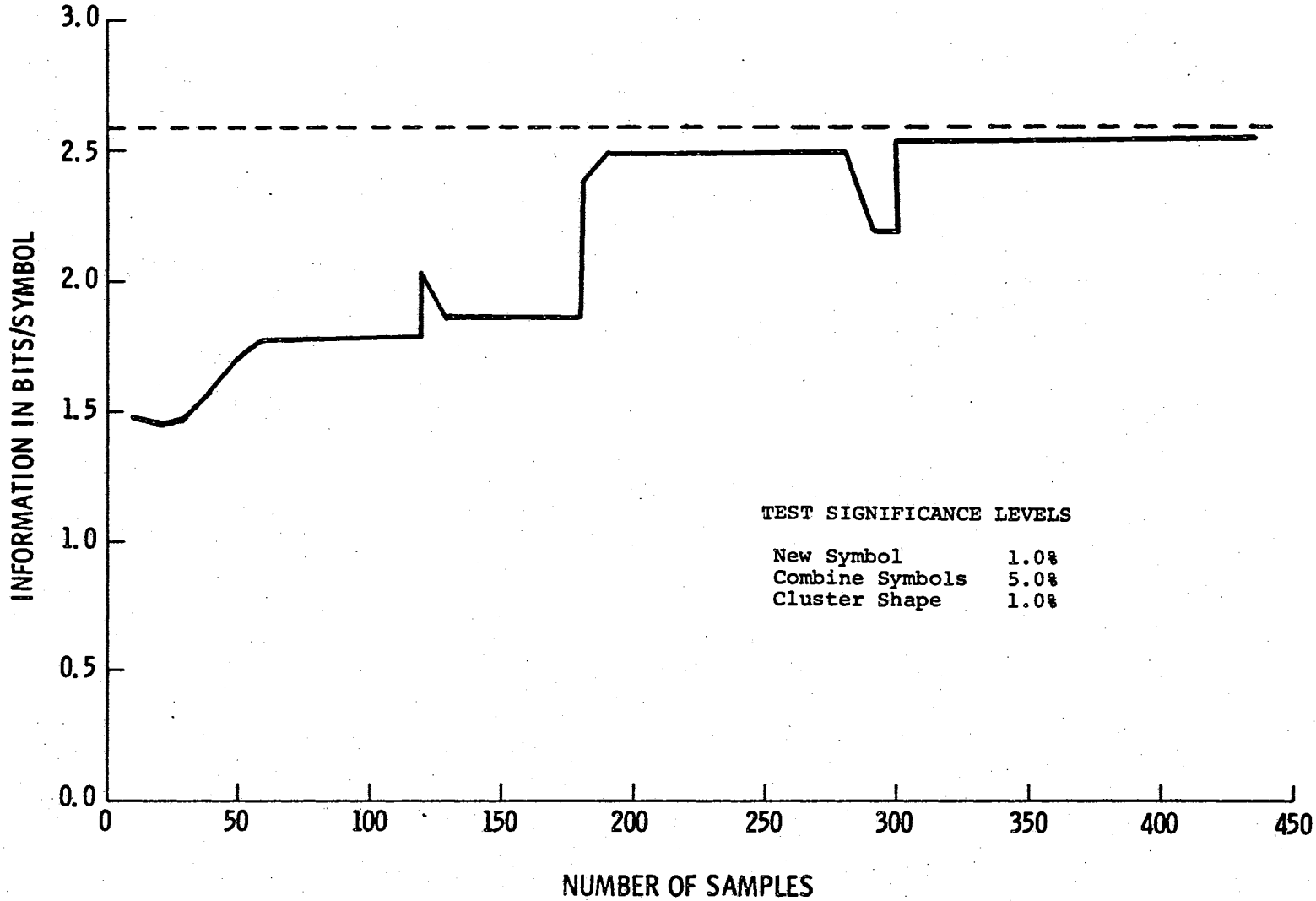

Figure 14 (Continued)

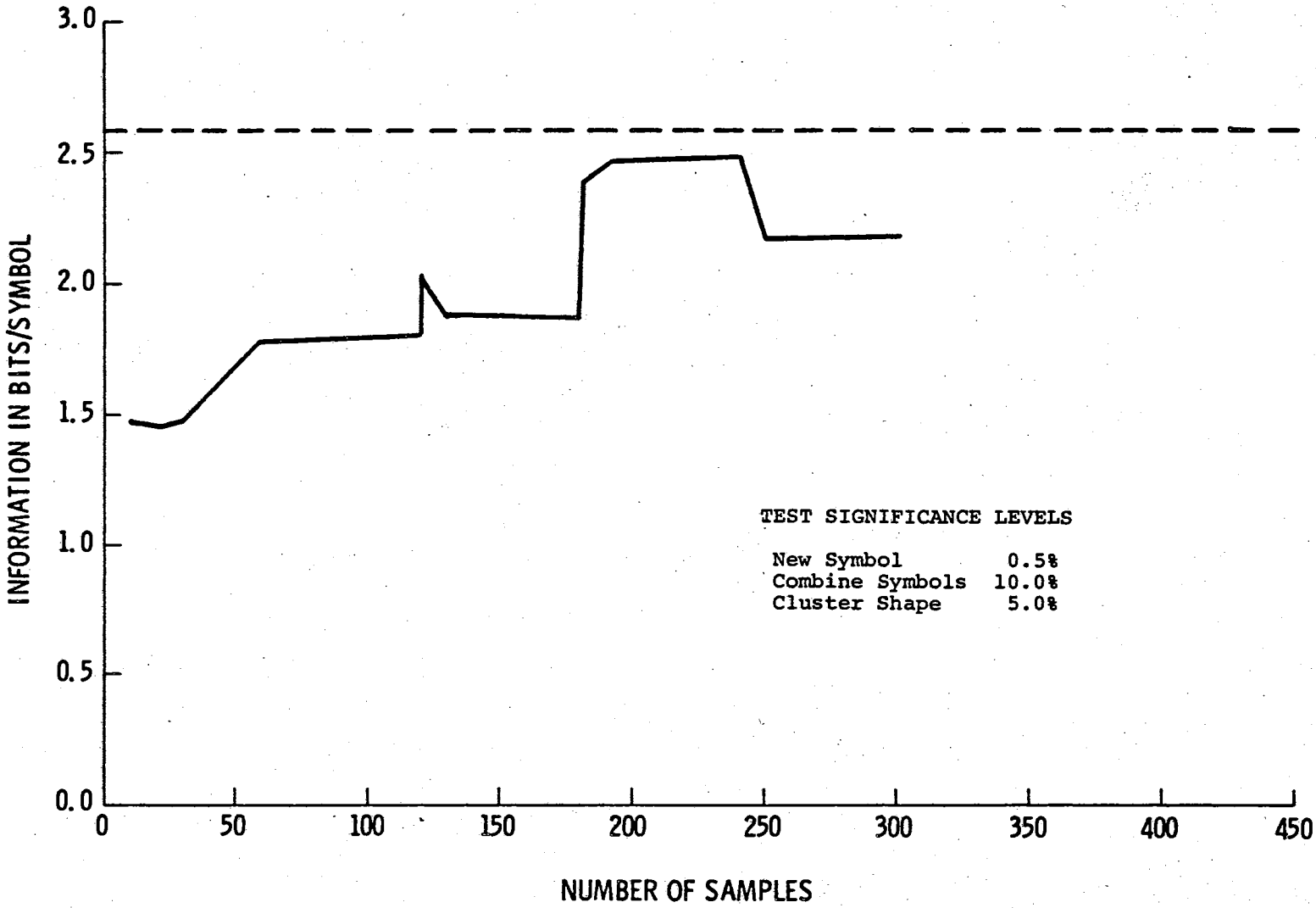

Figure 14 (Continued)

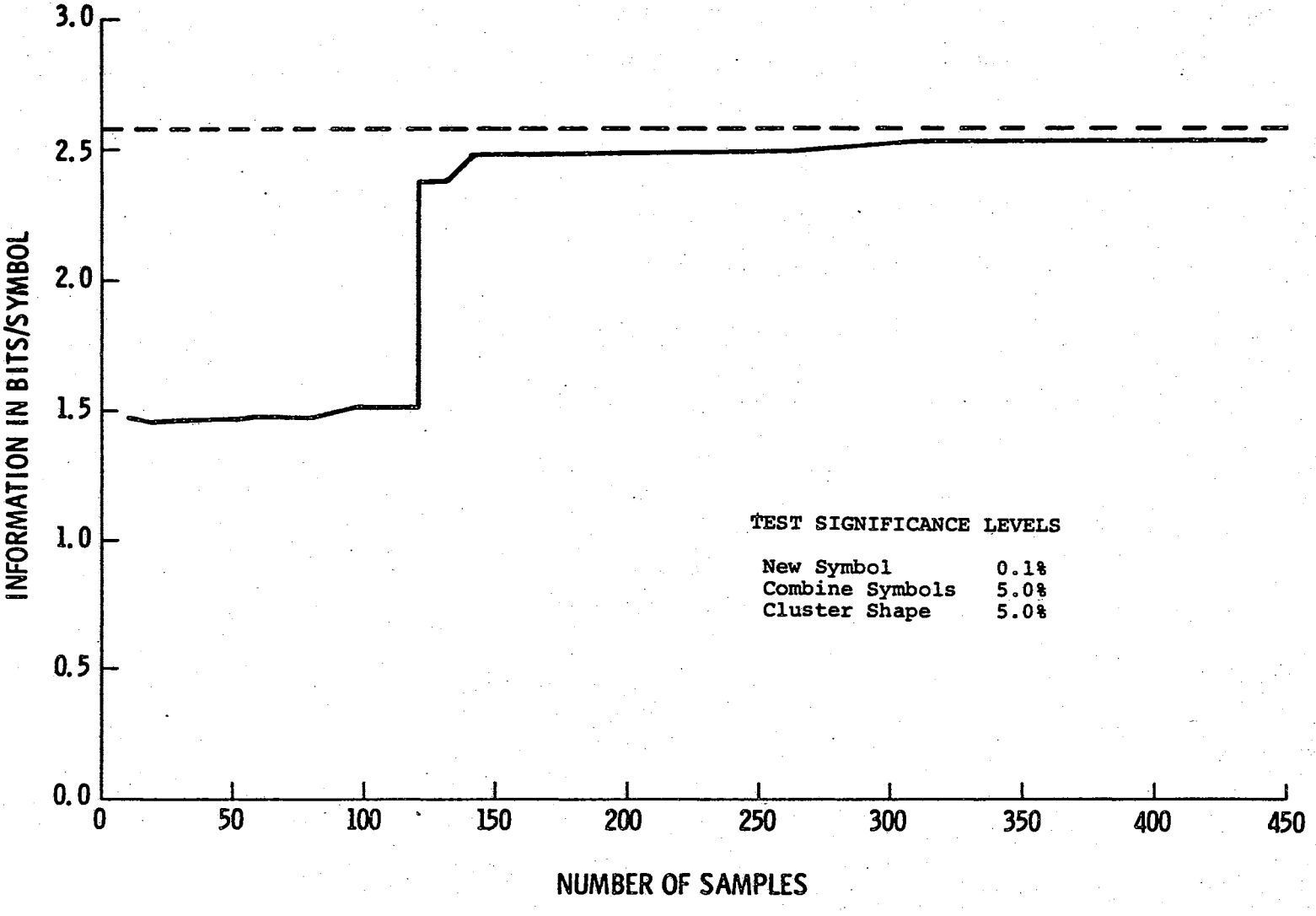

Figure 14 (Continued)

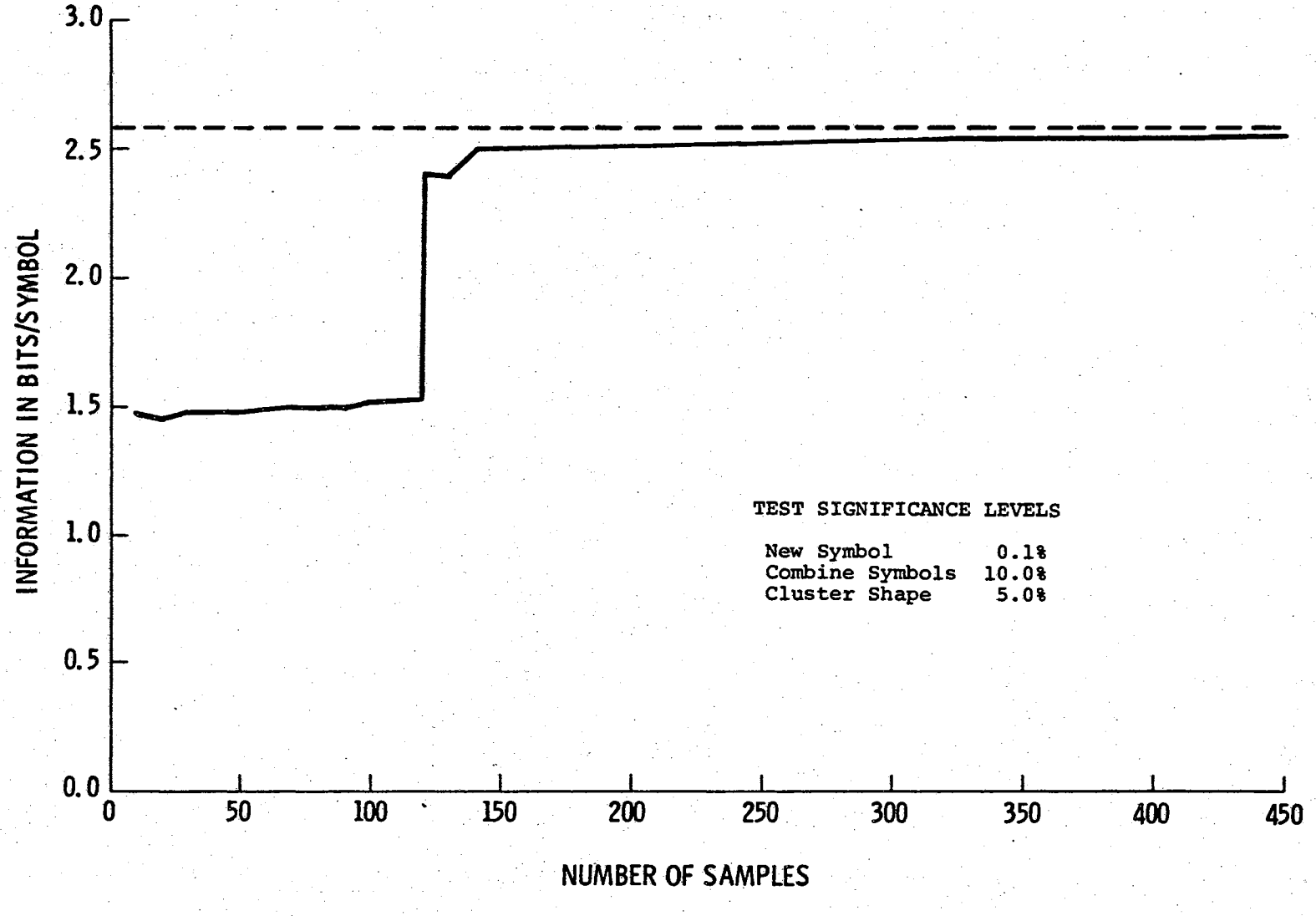

Figure 14 (Continued)

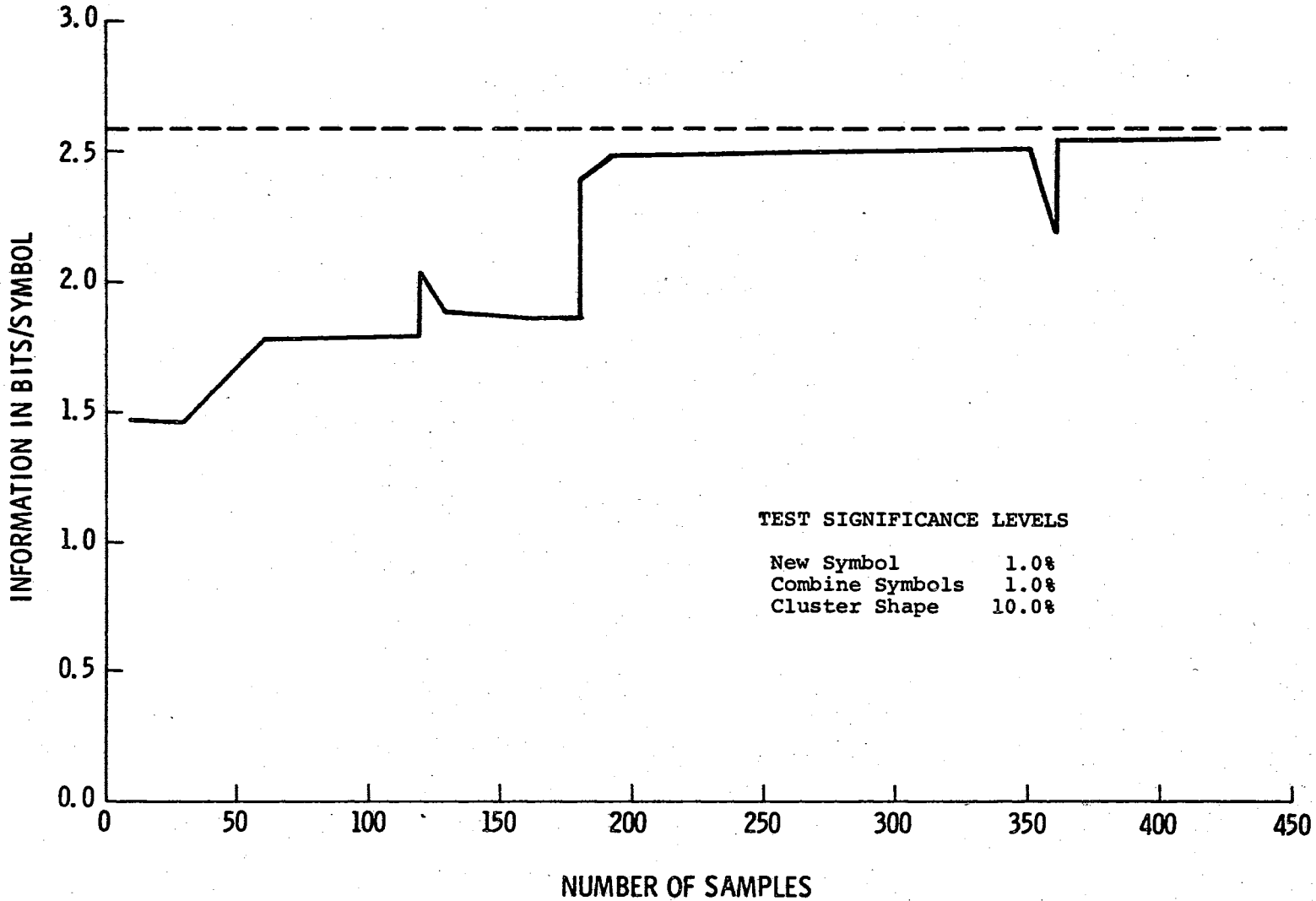

Figure 14 (Continued)

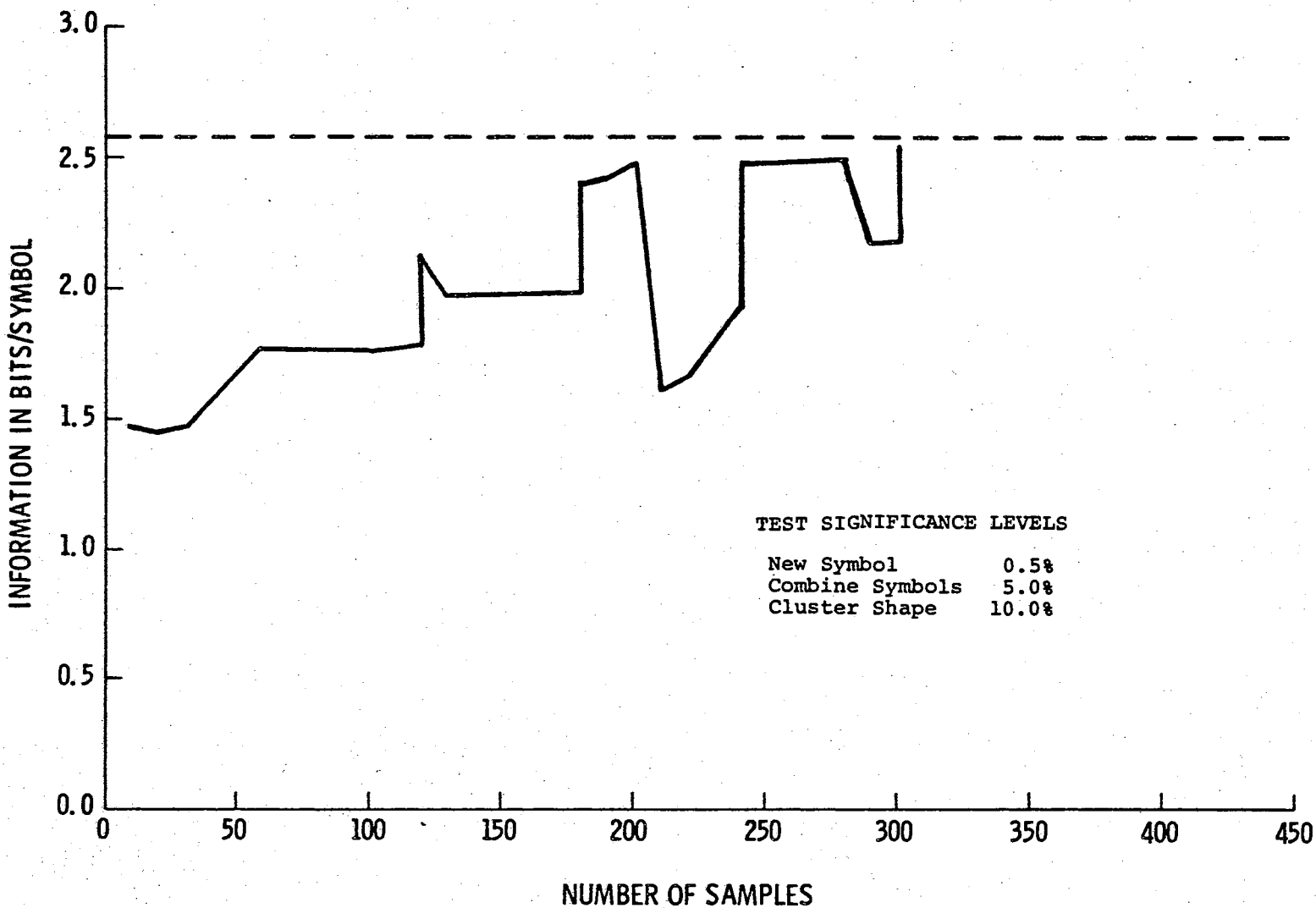

Figure 14 (Continued)

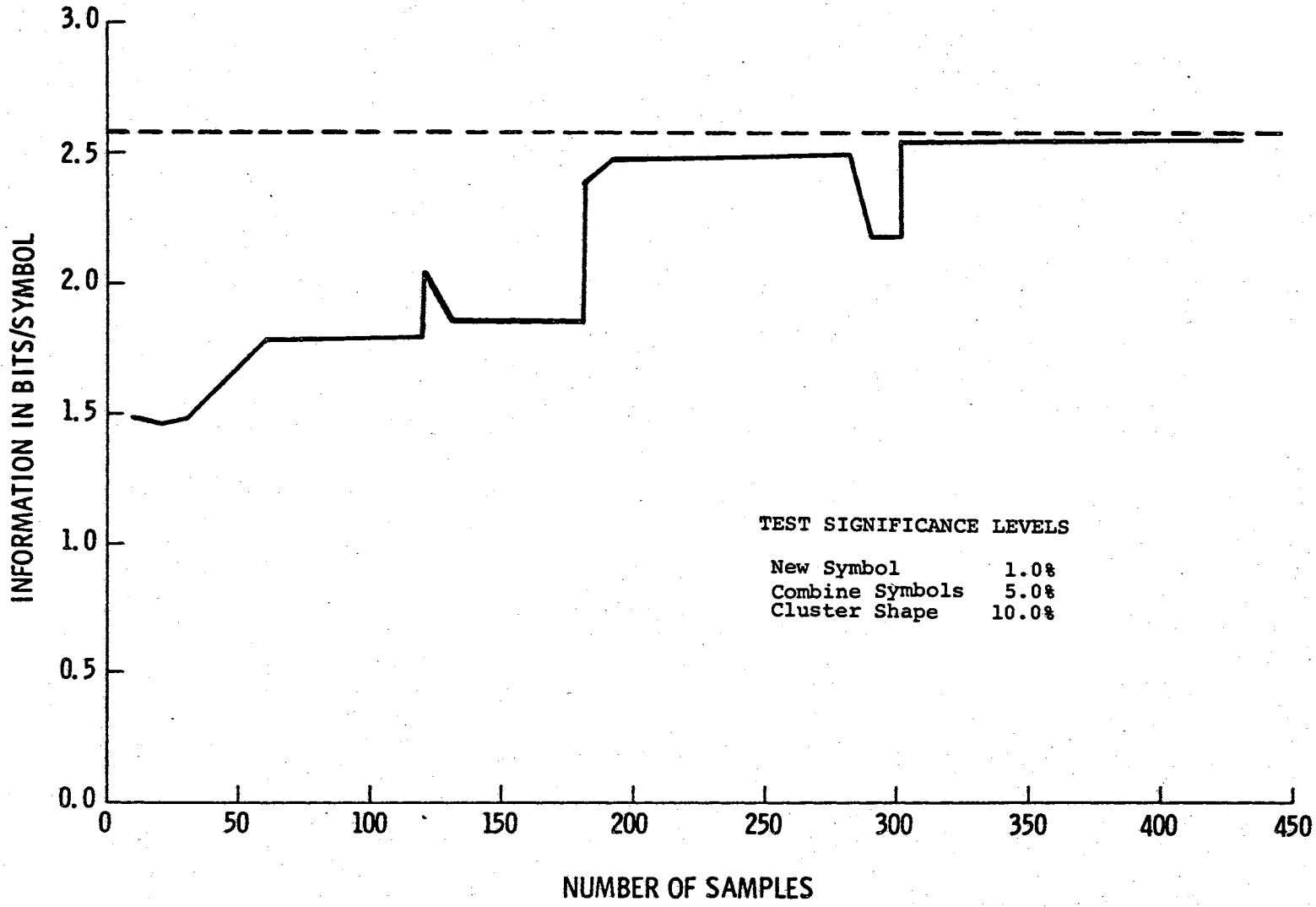

Figure 14 (Continued)

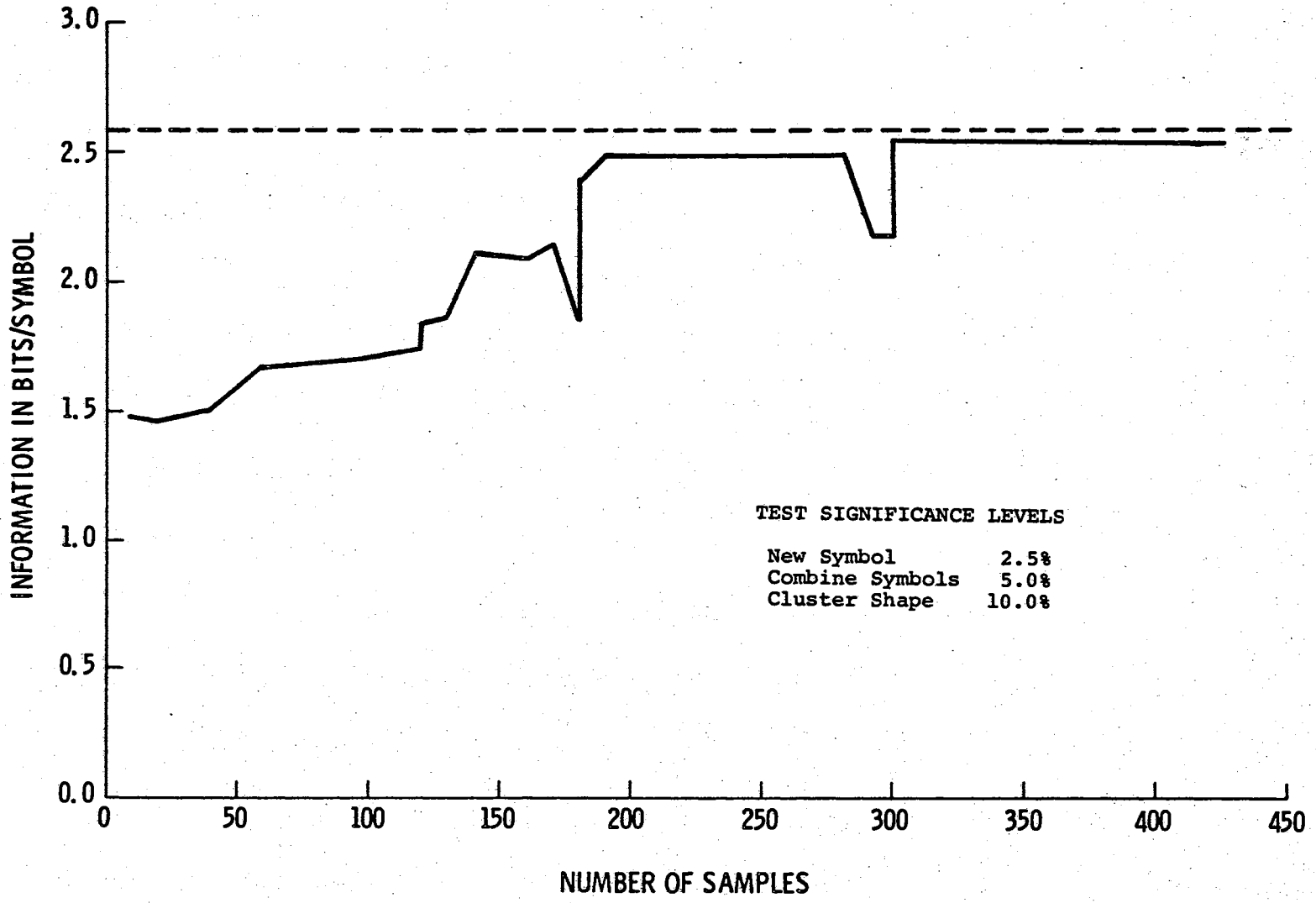

Figure 14 (Continued)

printed output of the program shows this to be true.

The new symbol test and symbol combining test are somewhat complementary in that any spurious new symbols generated through this medium should be removed promptly by subsequently combining them with an existing symbol. In order for this sequence of events to occur it is necessary first for the new symbol recognizing test to be sensitive to the noise induced signal disturbances. Yet, when the new symbol test significance level is greater than 1%, the effects of having too many receiver symbols begins to be noticeable in the learning curve. Generally this shows up as an unevenness in the learning curve such as appears in Figure 14a and 14b. If the receiver is fortunate enough to correctly postulate the new symbols, then real learning can take place here. Inspection of the printed output reveals, however, that in most cases the new symbols are spurious representations of one of the already accepted receiver symbols. Soon the symbol combining test eliminates the extra symbols. It seems that readiness of the receiver to declare new symbols without being too sure of itself tends to create more confusion in the learning process than it is worth. For the sort of data used in this example, the receiver learns most smoothly when the new symbol test has significance level in the neighborhood of 0.1%.

The main purpose of the symbol combining test is to eliminate spurious symbols generated by the new symbol test. As such, its significance level should not be critical as long as it is able to do its job without eliminating any true representations of the transmitter symbols. For the range of values investigated here, from 1% to 10%, the test seems to function satisfactorily. It is believed that somewhat larger levels would result in the unnecessary combining of symbols, slowing

the rate of learning. In this case the receiver might be forced to extend its recognition scheme to include a greater amount of data in the form of a large number of new symbol features before the differentness of the symbols would be great enough to satisfy the symbol combining test. The recognition and learning processes would then be considerably less efficient.

Considering all of the trial runs that were made in this investigation for the many different sets of data and various receiver parameter values, it appears that the optimum level for the new symbol test is less than 1%. The test should indicate a received signal to represent a previously unrecognized symbol only when it is virtually certain that is the case. The optimum value for the symbol combining test significance level seems to be about 5%. The receiver's learning process is somewhat more tolerant to having too few receiver symbols than too many during the early learning stage. The significance level of the cluster shape test is the least critical of all and the system seems to work well with this parameter at about 10%. Increasing the sensitivity of the shape test beyond the 10% significance level resulted in little change in performance except that the receiver became less efficient as it wasted time looking for new symbols in the data where there were none to be found.

Figure 15 shows the learning curve of one of the most efficient runs using the transmitter symbol set of Figure 12. The receiver was able to extract enough information from the first 120 signals to form good estimates of each of the transmitter symbols and to recognize all of the signals correctly. In more ordinary situations the receiver required about 50% more signals to make the same estimates. The rapidity

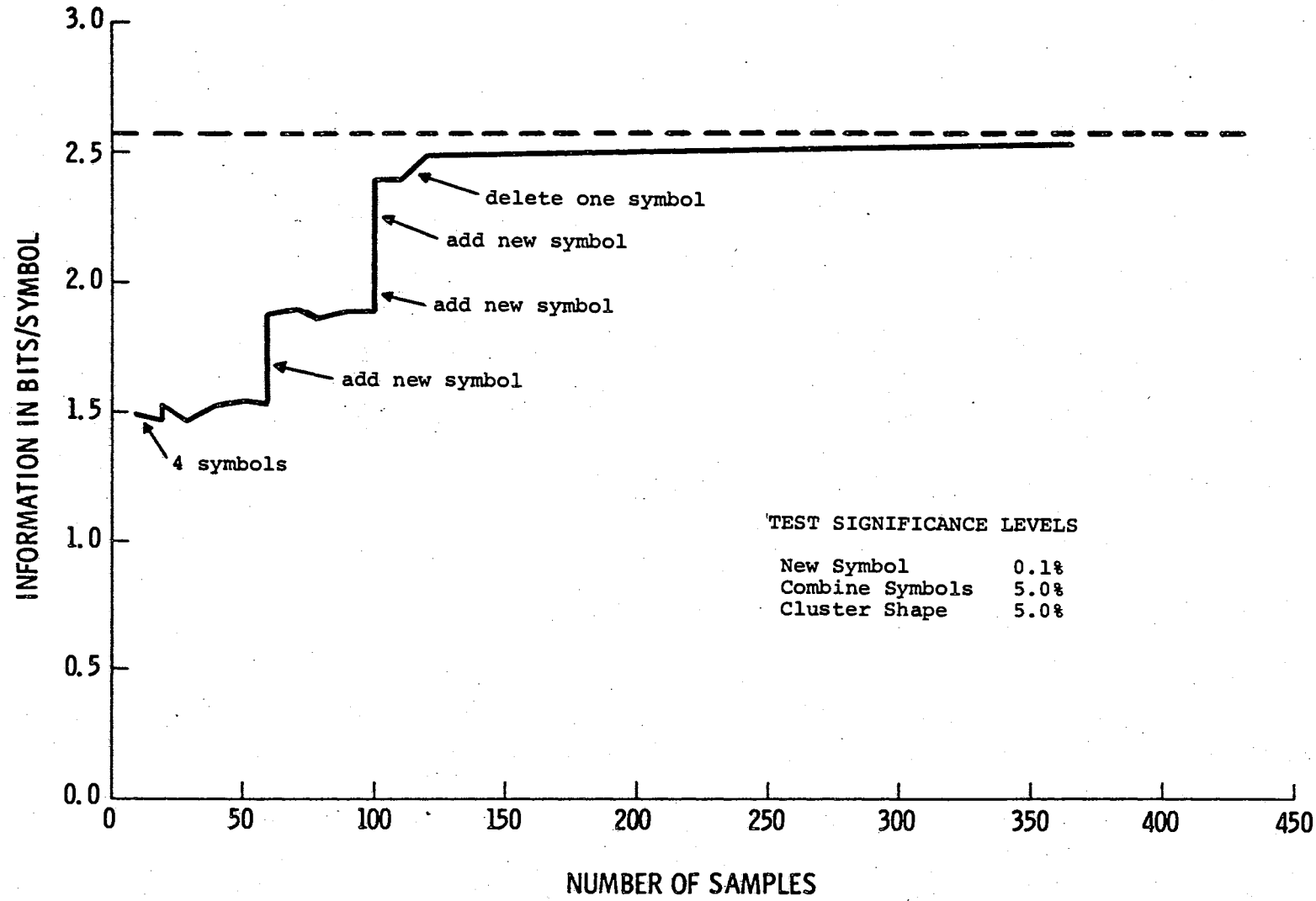

Figure 15. Example of Efficient Learning

of learning on this run seems to be the result of a fortuitous combination of receiver parameter settings and a "good" sequence of signals.

Interactions Between Functions Performed by the Receiver

Some further comments about the learning ability of the receiver may be in order at this point. There seem to be some subtle interactions between the various tests which cause symbols to be added and deleted at the receiver. Quite often a seemingly incorrect decision in one of the tests ultimately results in an improvement in overall performance when the decision is partially reversed by one of the other tests. It appears that the overall learning process is able to progress not only by its correct decisions but also by its mistakes. Certainly there is no reason to believe that this is impossible in a very general learning situation. Whether or not this takes place in the very simple model used here is perhaps a matter for personal interpretation.

For instance, consider the example shown in Figure 13. It is interesting to take note of the system's operation in reference to the number of symbols it proposes to receive. When the receiver has processed about 210 signals it deletes one symbol from its set leaving it with only five, one less than the number being transmitted. Almost immediately it adds a new symbol. Here attention is called to the performance improvement occurring as a result of the symbol deleting and adding perturbations in the recognition process. Before the disturbance, the system had settled down to a uniform level of performance. The improvement in information flow with corresponding reduction in number of errors occurs only because the recognition scheme stabilizes at an improved position after the disturbance. It is unlikely that the

system could have made a smooth transition between the before and after disturbance situations both because increasing the number of dimensions used in the symbol classifying process causes a complex change in the recognition scheme and because errors in the recognition process tend to perpertuate themselves and cause similar new errors to occur. It is only by upsetting the recognition scheme enough to mask the short-term effects of the past errors that the system is able to improve itself.

Here it is contended that the so-called disturbance to the recognition process is, in fact, a fundamental part of the hypothesis generating procedure employed by the learning receiver. The generation of new symbols for recognition by the receiver is not done randomly, but is designed to satisfy the requirements of the receiver based on its concept of self-satisfaction with its performance. The new symbols themselves are modeled by the data which started the search for new symbols through the receiver's handling of that particular data.

Generally, there is no attempt to alter the recognition of symbols with which the receiver is satisfied, although there is no assurance that such will be the case. The reprocessing and reorganization of the data in light of the new hypotheses generated within the receiver may lead to considerably different recognition of some or all the data even when the new hypothesis directly affects only a small part of the recognition scheme. In the example just considered, the generation of the new symbols for recognition by the receiver results in a reduction in the number of misclassified signals among all the symbols, even those not directly affected by the changes in the receiver symbol set. In this case the receiver learns to classify all of the signals properly when the last new symbol is proposed after about 275 signals have been

processed even though not all the previously made errors concerned that symbol.

It is important to realize here that the ability of the receiver to perform at this level comes more from the generation of the proper receiver symbol set than from the gradual accumulation of data one signal at a time upon which the estimates of the transmitter symbols are based. It is quite true that the information for estimating the transmitter symbols is contained in the body of received data but it takes the sudden proposing of new symbols to actually make use of that data. The data from the first 250 signals probably contains enough information on the six transmitter symbols to form very good estimates, but those estimates are not formed correctly by the receiver until it "discovers" a better way to process the data.

No doubt it would be stretching this point considerably to claim that the learning receiver has the faculty of "insight" for the formation of its new symbols. On the other hand, it certainly does not search completely at random either. In one respect the receiver is decision-directed in the formation of symbol estimates, yet it possesses the capability of broadly reviewing its past performance and attempting to improve on it.

It is during the review of past performance that the receiver may search for new discriminants upon which to base its decision rule for signals with whose recognition it is not completely satisfied. By considering additional information in the form of a higher dimensional signal vector the receiver may implement recognition schemes yielding considerably different results which, in turn, may motivate further searches for additional information on the part of the receiver.

Generally, there is no reason to believe that there should be any way to proceed directly to the final decision scheme without passing through the intermediate learning and recognition steps. The decision scheme implemented at each step certainly need not have any particular relationship to that of the previous step. The extreme non-linearity in the system's progress introduced by the addition of new information is perhaps the major source of difficulty in analyzing and predicting the system's behavior. This is perhaps the major difference between a true learning system and one which merely uses received data to modify the parameters of a fixed structure. In a truly adaptive learning system the fundamental structure may be changed to accomodate the data. Once again it would be stretching the point to claim that the model demonstrates this ability in a completely general fashion, yet there is a hint of this ability in the changing structure of the decision scheme which accommodates itself to the structure of the data.

It is the need for satisfaction with its own performance which is the foundation of the receiver's ability to learn. If the receiver is completely satisfied by its present performance, then no new decision schemes will be investigated and the learning of new symbols will stop. Under such circumstances it is still possible for the receiver to utilize some information in the data to refine its estimates of the transmitter symbols and so improve the recognition process slightly, but if the fundamental decision scheme is not correct at this point, at least insofar as the number of symbols to be recognized and their important features, the receiver will never be able to improve itself significantly.

The Effects of the Data on Learning in the Model

Thus far all attention has been devoted to the receiver in the communication system model and how it handles a given set of data. Certainly another area of importance is the data itself or, more specifically in the context of the simulation, just what properties of the transmitter symbols and the sequence in which they are sent enable and aid the learning receiver to perform its task.

As has already been noted, the short term learning progress of the receiver is affected strongly by the sequence in which the symbols are selected for transmission. It seems unlikely that this would have any significant effect on the long-term or average learning capabilities of such a system. This supposition is borne out by the small amount of data taken with this in mind. For example, Figure 16 shows two learning curves for different sequences of the same transmitter symbols. Evidently the receiver learns the transmitted symbols in a different way each time but the overall result is the same.

Certainly one would expect the "differentness" of the transmitted symbols to have an effect on how the receiver discovers new symbols in the data. Likewise, the geometrical arrangement and dimensionality of the vectors representing the transmitter symbol set could be important.

Figure 17 shows a typical learning curve for the receiver when it was presented with data generated by the transmission of six symbols equidistant from each other in a 5-dimensional signal space. The Euclidean distance between each pair of symbols was 4.95 times the noise standard deviation thus meeting the general requirement for "differentness" and permitting recognition with a very low error rate by the best decision scheme. The receiver was programmed to begin by

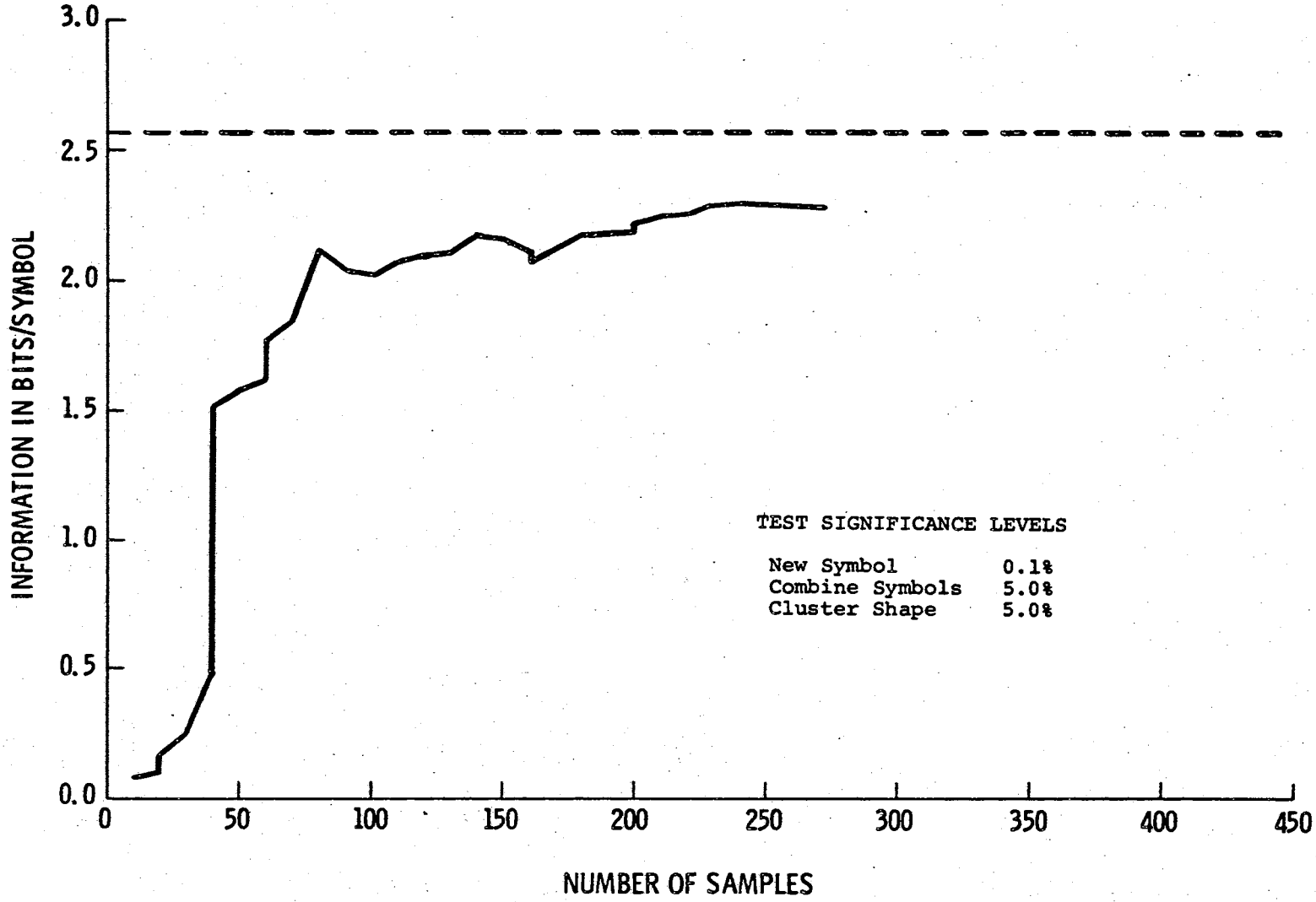

Figure 16. Learning Curves Showing Effect of Change in Sequence of Symbols

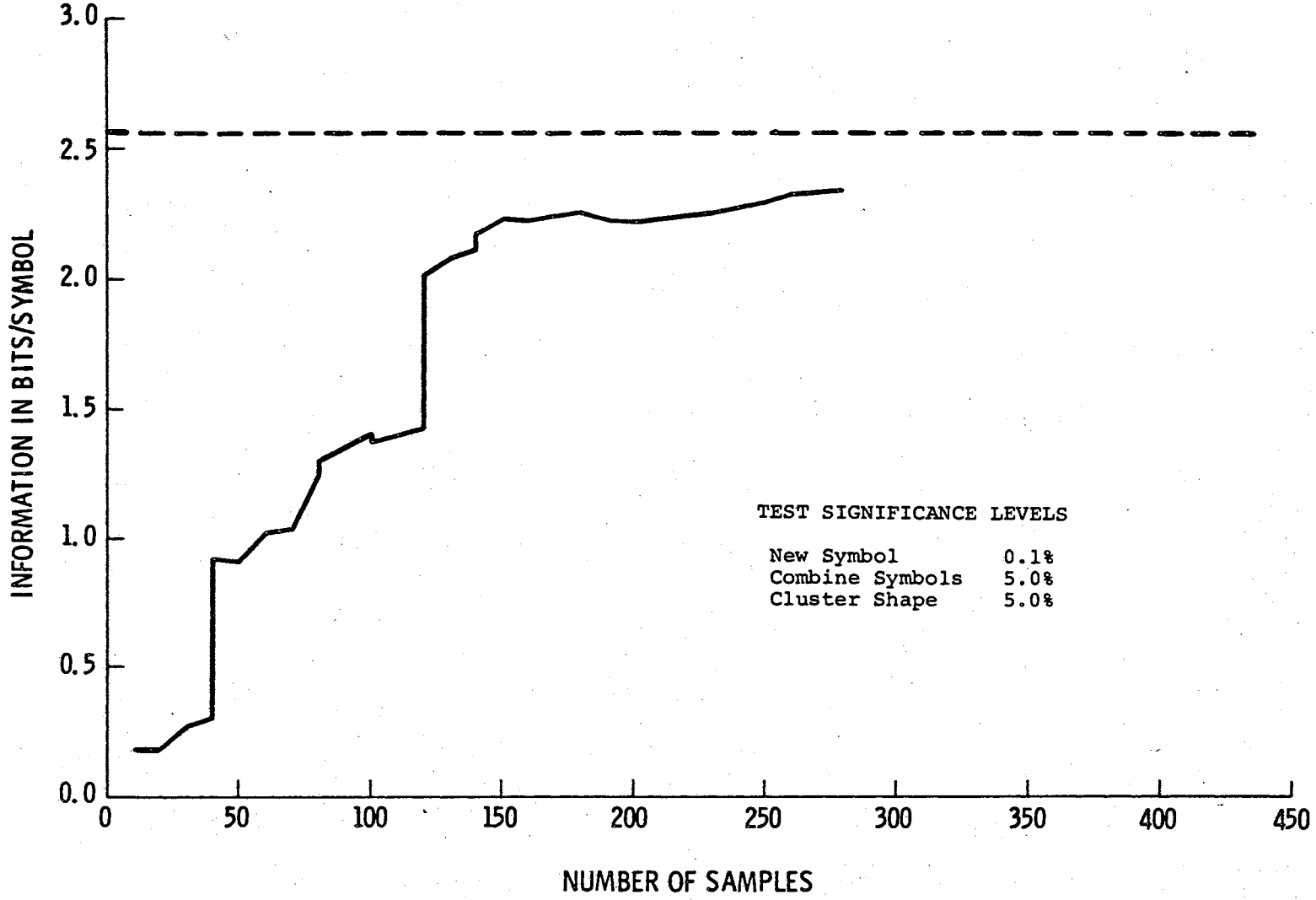

Figure 16 (Continued)

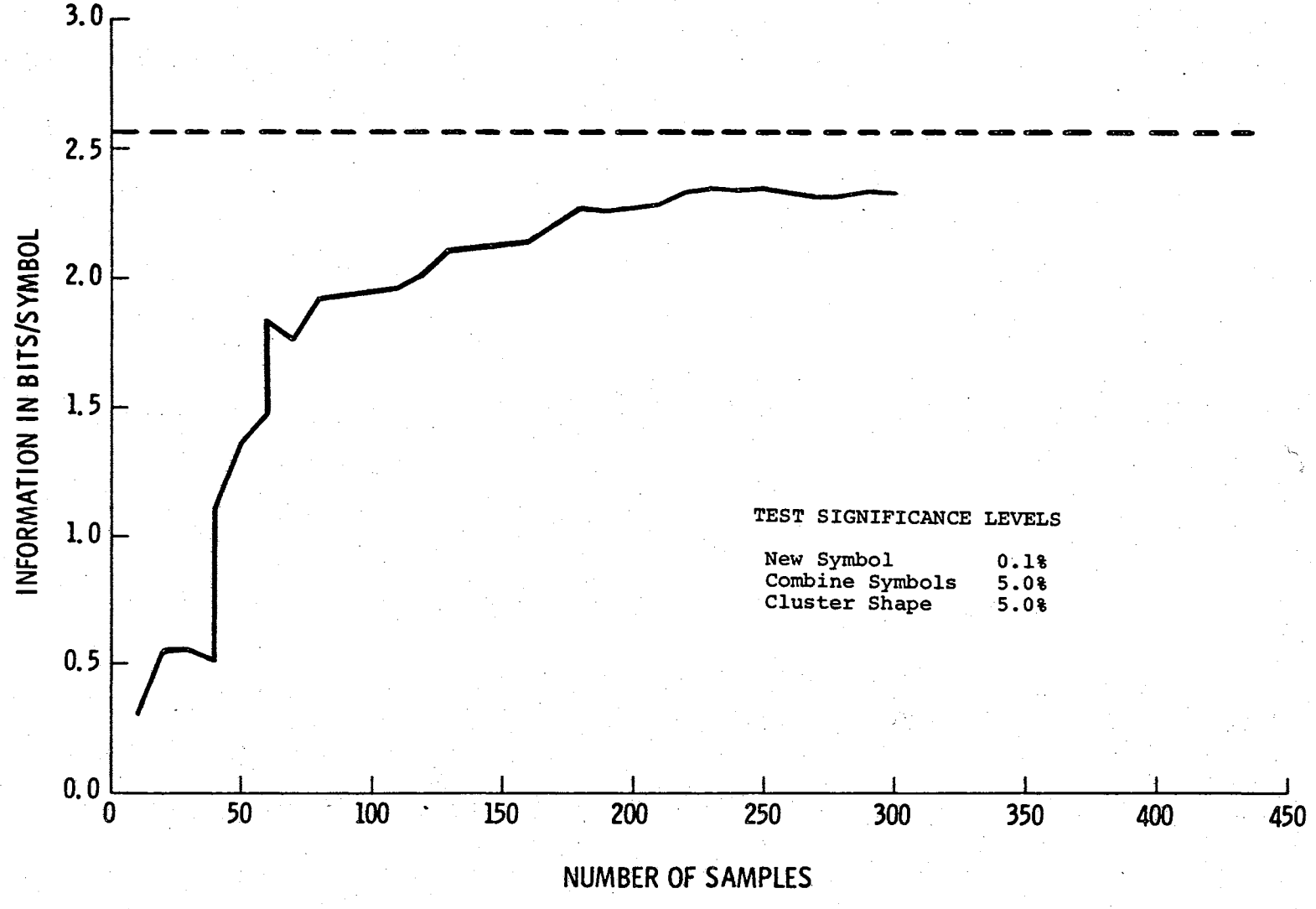

Figure 17. Example of Higher Dimensional Learning

considering all five dimensions of the received signal vectors, thereby utilizing all of the available information.

It is immediately noticeable in Figure 17 that the learning process as evidenced by the information flow curve seems to progress in a smoother and more orderly fashion than for the lower-dimensional cases cited previously. The more orderly arrangement of the transmitted symbols may well have something to do with this but experience with other higher dimensional data indicates that there are other factors involved. In the first place, there is a greater amount of learning necessary to achieve a certain information flow in the higher dimensional situations. That is, more numbers must be learned by the receiver in order to implement the recognition scheme at a desired level of "goodness" in higher numbers of dimensions. Thus, learning might be thought to occur more slowly although there is also more information available to the receiver from which to learn in the form of the higher dimensional signal vectors. Secondly, the distribution of errors or inaccuracy in the received information may be more evenly distributed because of the larger number of noise components on the received signal. Certainly the phenomenon of "sphere hardening" is taking place. The term sphere hardening is applied to the phenomenon where the length of the normalized noise vector added to the signal vector has less and less proportional variance as the dimensionality of the signal space increases. This means that the received signal vectors will tend more and more to lie on the surfaces of hyperspheres as the number of dimensions increases. Each hypersphere would be centered on the true symbol representation. It is not clear just how these various factors affect the process of learning as modeled here but it is evident that the dimensionality of

the signal space is itself a factor in the learning process.

If there is any conclusion to be drawn from these observations it seems to be that learning takes place more smoothly and predictably when information is received in the form of a large number of measurements each containing a small part of the total available information. On the other hand, it is more difficult to process the many-dimensional data so the computational efficiency of the learning process is reduced, at least when simulated on a digital computer and programmed in the manner of this investigation.

CHAPTER V

SUMMARY AND CONCLUSIONS

Summary

The phenomenon of hypothesis generation (generation of new ideas) in a learning system has been investigated. This effort was motivated by the general feeling that the generation of hypotheses is an essential part of the learning process and a part which is particularly poorly understood from a technical standpoint at the present time. Attention was focused on efficient ways of generating new hypotheses and how the learning process, i.e., the generation of "good" hypotheses, was influenced by data presented to the learning system. It was demonstrated that these good new hypotheses could be generated rather efficiently by a consideration of the data rather than by a random process of trial and elimination.

There was a distinction made between generating completely new hypotheses and modifying old ones. In the context of present-day learning systems, attention centered on machines which possessed the ability to change their basic structure rather than only certain parameters of a fixed structure.

The general approach taken in this investigation has been to model the learning system as a decision-making device where the various alternatives considered by the system correspond to hypotheses which have been learned in the course of its experiences. Included in the concept

of a hypothesis is not only a possible decision but also some rule for how that decision is to be reached. Taken collectively, the set of hypotheses determine both the set of possible decisions and how those decisions are reached from raw data presented to the system.

In the communication system model proposed here, the hypotheses correspond to symbols proposed for reception by the receiver. The recognition scheme based on these proposed symbols determines the decision rule. The geometrical interpretation given to the hypothesis generation and recognition processes tends to aid the understanding of the receiver's operation and particularly demonstrates at least one way in which new hypotheses may be generated in a very efficient manner by the learning receiver.

The computer simulation of the model demonstrated the feasibility of such a model of a learning system and gave some indication of the broad range of conditions under which it could function effectively. Such a simulation was particularly useful in this study because of the difficulty in finding tractable methods for analyzing the behavior of this complex system.

The results obtained from the simulation were much more brief than had originally been intended. Because of the large amount of computer time required by the simulation program, only about 30 useful sets of data were available for analysis. Perhaps this was just as well since nothing completely unexpected is evident in the results which were obtained in this manner.

The most obvious results are the demonstration that the model works as the theory predicts and that it is not particularly sensitive to its own internal operational parameters. The generation of new hypotheses

based on data which is not consistent with existing hypotheses seems to be a valid mechanism for use in such a learning system. The fact that the method is not sensitive to the system parameters indicates that the principle may be useful in practical situations.

Comparison of Results With the Work of Others

In an attempt to gain some comprehension of the significance and relevance of the learning system model which has been proposed and investigated in this paper, one is inevitably drawn to make comparisons between this system and those which have been proposed by others. The general problem of teacherless learning in pattern recognition and classification systems has been treated in the literature rather extensively since about 1962 and a number of practical solutions to the problem have been proposed and investigated (14,15,16,17). Almost invariably these solutions have required knowledge of the number of different classes to be learned and recognized. All of the solutions for which substantial mathematical treatment has been offered have required this knowledge.

The model proposed and investigated in this study requires a considerably less restrictive kind of information. As indicated in the description of the model, the shape (number of modes) and variance of the noise distribution can provide enough information to insure success of the method. It is believed that the proposed model could function almost as effectively with even less a priori information than this since the model is directed more by the differences it discovers in its observations than by the comparison of those observations with some absolute standard.

MacQueen (10) describes a system for the classification and analysis of multivariate observations which is basically similar to part of the proposed model. The number of classes or categories recognized may be varied to suit the data using a set of fixed distance thresholds. The model proposed herein goes several steps further by allowing the thresholds to vary in a manner intended to give some statistical substance to the decisions for creating and exterminating categories.

Perhaps the major difference between the model investigated here and all others which have been reported to date is the ability of this model to reconsider the data and to search for new criteria for classification by varying the dimensionality of the signal space. There are two different but interrelated aspects to this facet of the model's operation. First, by measuring various properties, e.g., generalized variance, of large amounts of data, the parts of the data which may require more careful analysis are pointed out. Broadly stated, the data which is not "explained" by the present state of knowledge is marked for further investigation. This corresponds to the general scientific method in present-day research. The second part of the learning process, considering more information in the form of higher dimensional data upon which to base a new decision scheme, seems to be a natural extension of the pattern recognition problem except that in this particular application the suspect data itself is used to form the new decision scheme. A new hypothesis is proposed which, in fact, does explain the observed data. This, too, corresponds to the scientific method in research so there is reason to feel that the proposed algorithms are not inconsistent with the real world.

Concluding Comments

There are a number of general comments about the model and its study which, if not obvious at the outset of the investigation, rapidly became apparent as work progressed. First of all, the model is an attempt to explain what is probably a very complicated learning phenomenon by rather simple means. In the introduction it was stated that the reason present-day learning machines are not more successful is because the underlying phenomena are not understood. That was generally found to be the limiting factor in this study. No doubt, a larger and faster computer would have been of some help but, even so, the computer simulation seems to have been limited more by the theory than by the available hardware. A faster computer would have produced such a mountain of data that its analysis would have been as large a problem as the original investigation. If anything could have aided the study, it would have been a simpler theory to explain a better-understood learning phenomenon. Whether or not this is possible remains to be seen.

There is an interesting interpretation which may be given to the study results in view of the learning theory investigated and the limited amount of data available to test that theory. The study was initiated partly because of some pre-conceived ideas about the aspects of learning which were to be investigated and the results which were expected to be found. It was axiomatic, according to the theory, that preliminary results of the simulation could be and, in fact, would be explained by these pre-conceived ideas. The theory states that any discrepancy between these ideas and the "truth" should have become increasingly apparent as more and more data was obtained. Unfortunately for the exposition of this aspect of the theory, the amount of data

obtained was rather small in comparison to the complexity of the phenomenon investigated and so no significant discrepancies were discovered. There was very little in the line of unexplainable behavior by the system model to lead the investigation.

There is perhaps at least one exception to the above statement. The model's reaction to high-dimensional data has been noted as not being adequately explained by the theory. It is possible and, in fact, probable that some more detailed theory can be advanced to explain this phenomenon.

The use of heuristic methods in the model deserves some further comment at this point. The obvious reason for the use of such techniques is that the theory has not been expressed in strict enough mathematical terms to permit analysis and simulation on a more fundamental level. Whether this will ever be possible and/or worthwhile cannot be determined at this time. Certainly it is more desirable at the present time to gain an understanding of the overall learning process than to be concerned with the details, thus the justification for the loose mathematical framework.

In the introduction it was stated that because of the exploratory nature of the study and the limited amount of relevant data, the effort needed to build a rigorous theory may not be justified at the present time. The simulation results tell us that we need a reasonably large body of relevant data to point the way toward a precise and useful formulation of the theory. Certainly many minor variations of the theory could be supported by the available data. Clarification of many small points may be discovered and tested in the course of future work.

Recommendations for Further Study

At this stage of any investigation of this type it invariably appears that there is more work ahead than behind. This study is no exception. The theory could be expanded and investigated further in almost any direction. There are several areas touched on by this investigation in which some groundwork has been done to point out the limits of present-day knowledge. (1) There is the question of how much a priori information is necessary to insure convergence of the learning system to the "correct" solution. The work of Teicher (18) and Patrick and Hancock (19) on the identifiability of finite mixtures gives conditions under which convergence of a particular type system is certain but little is known about the sensitivity of more general learning systems to this a priori information. In particular, the question of interaction between a priori information and information available in the data but perhaps unused by the learning system is almost completely unanswered. (2) There is a need for a more comprehensive measure of the amount of learning or state of knowledge in such systems. The information flow measure used in this study can be formulated in several different ways because of various interpretations given to the empirical (relative frequency) probabilities and the subjective probabilities (credibilities). In any event, this particular measure is useful probably only in the communication system model. (3) The effect of finite memory capacity and time span has been neglected in this study. The addition of a finite memory model appears to be a straightforward extension but, as has been noted earlier, the effects might bring unexpected results.

SELECTED BIBLIOGRAPHY

1. Newell, A., J. C. Shaw, and H. Simon. "Chess Playing Programs and the Problem of Complexity." IBM Journal of Research and Development, Vol. 2 (1958), 320-335.

2. Samuel, A. L. "Some Studies in Machine Learning Using the Game of Checkers." IBM Journal of Research and Development, Vol. 3 (1959), 211-229.

3. Watanabe, S. "Information-Theoretical Aspects of Inductive and Deductive Inferrence." IBM Journal of Research and Development, Vol. 4 (1960), 208-321.

4. Minsky, M. "Steps Toward Artificial Intelligence." Proc. IRE., Vol. 49 (1961), 8-30.

5. Bakan, D. "Learning and the Principle of Inverse Probability." Psychological Review, Vol. 60 (1953), 360-370.

6. Basore, B. L. and W. D. Wood. "A Model for Communication With Learning." ASTIA Doc. No. AD-242535. Albuquerque, New Mexico: The Dikewood Corporation, May 31, 1960.

7. Schwartz, M., W. R. Bennett, and S. Stein. Communication Systems and Techniques. New York: McGraw-Hill, 1966.

8. Hill, B. N. "Information for Estimating the Proportions in Mixtures of Exponential and Normal Distributions." J. American Statistical Association, Vol. 58 (1963), 918-932.

9. Nagy, G. "State of the Art in Pattern Recognition." Proc. IEEE., Vol. 56 (1968), 836-862.

10. MacQueen, J. "Some Methods for Classification and Analysis of Multivariate Observations." Proc. 5th Berkeley Symp. on Statistics and Probability. Berkeley, California: University of California Press, 1967, 281-297.

11. Bishop, D. J. "A Test for the Homogeneity of Variances and Covariances." Biometrika, Vol. 31 (1948), 31-55.

12. Shannon, C. E. "A Mathematical Theory of Communication." Bell Sys. Tech. Journal, Vol. 27 (1948), 379-423, 623-656.

13. Fano, R. M. Transmission of Information. Cambridge, Mass.: M.I.T. Press, 1961.

14. Nagy, G. and G. L. Shelton, Jr. "Self-Corrective Character Recognition System." IEEE Trans. on Information Theory, Vol. IT-12 (1966), 215-222.

15. Fralick, Stanley C. "Learning to Recognize Patterns Without a Teacher." IEEE Trans. on Information Theory, Vol. IT-13 (1967), 57-64.

16. Nagy, George. "State of the Art in Pattern Recognition." Proc. IEEE., Vol. 56 (1968), 836-862.

17. Ho, Yu-Chi, and Ashock K. Agrawala. "On Pattern Classification Algorithms, Introduction and Survey." Proc. IEEE., Vol. 56 (1968), 2101-2114.

18. Teicher, H. "Identifiability of Finite Mixtures." Ann. Math. Stat., Vol. 34 (1963), 1265-1269.

19. Patrick, E. A., and J. C. Hancock. "Nonsupervised Sequential Classification and Recognition of Patterns." IEEE Trans. on Information Theory, Vol. IT-12 (1966), 362-372.

VITA

Charles Richard Reeves

Candidate for the Degree of

Doctor of Philosophy

Thesis: SOME NEW IDEAS AND TECHNIQUES FOR LEARNING SYSTEMS

Major Field: Electrical Engineering

Biographical:

Personal Data: Born in Jefferson City, Missouri, December 16, 1940, the son of Mr. and Mrs. Charles Victor Reeves.

Education: Graduated from Helias High School, Jefferson City, Missouri, May, 1958; received the Bachelor of Science degree from the University of Missouri at Rolla in 1962, with a major in Electrical Engineering; received the Master of Science degree from Oklahoma State University in 1966, as a National Science Foundation Trainee, with a major in Electrical Engineering; completed requirements for the Doctor of Philosophy degree at Oklahoma State University, as a National Aeronautics and Space Administration Trainee, in May, 1970.

Professional Experience: Electrical Engineer, Collins Radio Company, 1962-1964; joined Texas Instruments Incorporated as a Member of the Technical Staff, June, 1969.