

UNCLASSIFIED

AD NUMBER
AD464022
NEW LIMITATION CHANGE
TO Approved for public release, distribution unlimited
FROM Distribution authorized to U.S. Gov't. agencies and their contractors; Administrative/Operational Use; MAR 1965. Other requests shall be referred to Office of Naval Research, One Liberty Center, 875 North Randolph Street, Arlington, VA 22203-1995.
AUTHORITY
onr ltr 24 nov 1967

THIS PAGE IS UNCLASSIFIED

UNCLASSIFIED

AD_ 4 6 4 0 2 2

DEFENSE DOCUMENTATION CENTER

FOR

SCIENTIFIC AND TECHNICAL INFORMATION

CAMERON STATION ALEXANDRIA, VIRGINIA



UNCLASSIFIED

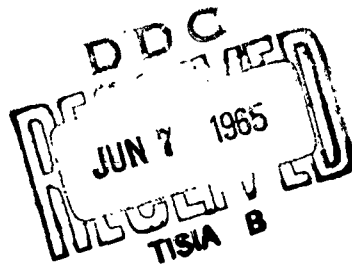
NOTICE: When government or other drawings, specifications or other data are used for any purpose other than in connection with a definitely related government procurement operation, the U. S. Government thereby incurs no responsibility, nor any obligation whatsoever; and the fact that the Government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

Learning to Recognize Patterns Without a Teacher

by
Stanley C. Fralick

March 1965

AVAILABLE COPY WILL NOT PERMIT
FULLY REPRODUCIBLE EDITION.
REPRODUCTION WILL BE MADE IF
REQUESTED BY USERS OF DDG.



Technical Report No. 6103-10

Prepared under

Office of Naval Research Contract

Nonr-225(83), NR 373 360

Jointly supported by the U.S. Army Signal Corps, the

U.S. Air Force, and the U.S. Navy

(Office of Naval Research)

SYSTEMS THEORY LABORATORY

STANFORD ELECTRONICS LABORATORIES

STANFORD UNIVERSITY • STANFORD, CALIFORNIA

464022

CATALOGED BY: DDG

AS AD NO. 464022

DDC AVAILABILITY NOTICE

Qualified requesters may obtain copies of this report from DDC.
Foreign announcement and dissemination of this report by DDC
is limited.

LEARNING TO RECOGNIZE PATTERNS WITHOUT A TEACHER

by

Stanley C. Fralick

March 1965

Reproduction in whole or in part
is permitted for any purpose of
the United States Government.

Technical Report No. 6103-10

Prepared under
Office of Naval Research Contract
Nonr-225(83), NR 373 360
Jointly supported by the U.S. Army Signal Corps,
the U.S. Air Force, and the U.S. Navy
(Office of Naval Research)

Systems Theory Laboratory
Stanford Electronics Laboratories
Stanford University Stanford, California

ABSTRACT

The techniques of decision theory are applied to the problem of constructing machines that improve their ability to recognize patterns by extracting pertinent information from a previously unclassified sequence of observations; such machines are said to learn without a teacher.

A general system solution is obtained which includes the solutions to the problems of learning without a teacher, learning with a teacher, and no learning. The solution has been extended to include problems in which the unknown parameter is time varying, as well as problems in which the probabilities of occurrence of the classes are unknown a priori and must be learned. The resulting systems are shown to be stable and to have performance which converges to the performance of systems which have a priori knowledge of the unknown parameters being learned. It has been demonstrated that for most cases either the optimum system, or a suboptimum system which performs within an arbitrarily small tolerance of the optimum system, is realizable in the sense that it requires a finite memory.

The techniques of this paper are applied to examples of learning problems in the communications, radar, and electromagnetic reconnaissance fields.

CONTENTS

	<u>Page</u>
I. INTRODUCTION	1
A. Classification Problems	1
B. Decision Machines Which Learn	3
C. Related Work	4
D. Organization, Approach, and Significant Results	6
II. DEVELOPMENT OF THE LEARNING SYSTEM	9
A. The Learning Problem Model (Binary Decisions)	9
B. An Exponentially Growing Solution	11
C. A Recursive Solution	13
D. Examples	17
1. Detection of a Signal of Unknown Amplitude	18
2. Detection of a Narrowband Signal of Unknown Frequency	20
E. Learning the A Priori Probabilities	25
F. Extension to the Multiple-Hypothesis Decision Problem	30
G. Summary of Chapter II	32
III. PERFORMANCE OF LEARNING SYSTEMS	33
A. Performance Measures	33
B. A Technique for Bounding the Performance	34
C. Example	39
D. Other Techniques To Obtain Performance Bounds	42
IV. LEARNING TIME-VARYING PARAMETERS	46
A. Models for the Time-Varying Parameter Problem	46
B. Solution to the Problem	48
1. Case 1, General-Random-Walk Time Dependence	49
2. Case 2, Binomial Time Dependence	54
C. Examples	58
1. The Fading-Channel Problem	58
2. Frequency-Hopping Signal Reconnaissance Problem	70
D. Summary of Chapter IV	72

	<u>Page</u>
V. SYSTEM REALIZABILITY	73
A. System Memory Capacity	73
B. Minimal Sufficient Statistics	74
C. Practical Considerations	78
D. Summary of Chapter V	81
VI. SOME PROPERTIES OF LEARNING SYSTEMS	83
A. System Stability	83
B. Convergence of the Continuous System	84
C. Convergence of the Quantized System	85
D. Relationship between Learning with a Teacher and Learning without a Teacher	88
E. Summary of Chapter VI	90
VII. SUMMARY OF RESULTS AND SUGGESTIONS FOR FUTURE WORK	91
A. Results	91
B. Problems for Additional Research	92
APPENDIX A. Evaluation of $P(B_k)$	93
APPENDIX B. Proof of Theorem 1	98
APPENDIX C. Proofs of Stability and Convergence	101
REFERENCES	114

ILLUSTRATIONS

<u>Figure</u>	<u>Page</u>
1 A system to compute $p(\theta \lambda_{k-1})$	15
2 A binary detector which learns without a teacher (sequential form)	16
3 A binary detector which learns without a teacher (parallel form)	17
4 A learning system for detection of signals of unknown amplitude	20
5 A computer for $\ell(X f_1)$	26
6 A learning receiver	26
7 A system designed to learn the a priori probabilities . .	29
8 A multiple-hypothesis machine	30
9 A multiple-hypothesis machine which learns without a teacher	32
10 An optimal learning system	35
11 Performance bounds for a learning system	41
12 Learning system for general-random-walk time dependence . .	51
13 Probability computer, case 2b	57
14 Probability computer, case 2c	57
15 Learning system, case 2c	58
16 The fading-channel model	59
17 Learning receiver for fading on-off keyed signals	64
18 Probability computer	66
19 Tapped-delay-line realization of time-varying linear filter	67
20 Learning receiver for fading frequency-shift keyed signals	70
21 Likelihood computer	72

SYMBOLS

a	the amplitude of a narrowband signal, a scalar random variable
A	the parameter of a Rayleigh distribution
$\{A_i\}$	a set of vector-valued unknown parameters
A_k	the event $P(\hat{\theta} \lambda_k) > P(\theta_i \lambda_k)$ for all $\theta_i \neq \hat{\theta}$ (see Chapter III)
$b(t)$	a known waveform of finite time-bandwidth product
B	a known column vector with the sample values of $b(t)$ as elements
B_k	the event $P(\hat{\theta} \lambda_k) \leq P(\theta_i \lambda_k)$ for some $\theta_i \neq \hat{\theta}$
c	an unknown scalar parameter; also used as an index to indicate "current value of"
C_i	a constant $= A^2 / (A^2 R_i^2 + 1)$
d	divergence (see Chapter III)
$d(\cdot)$	a decision rule
D	a constant $= 1/S_n(f_i)$
$E(f_i)$	a column vector with the m^{th} element $= \exp(j2\pi f_i m \Delta)$
$E\{\cdot\}$	the expectation of the quantity within braces
f, f_i	the frequency of a sinusoid
$f(\cdot)$	a function used to factor a statistic
g, g_k	the gain, or attenuation, of a randomly time-varying communication channel. The index indicates the value is to be taken at a particular time, kT .
$g(\cdot)$	a function used to factor a statistic
G	a complex constant used to represent the operation of a fading channel on a signal. The modulus of G is the channel gain and the argument of G is the channel phase shift.
$h(\cdot)$	a function used to factor a statistic

$h(t, \gamma)$	the response of a time-varying linear filter at time t to an impulse applied at time γ
H_i	the i^{th} hypothesis
i	an integral-valued index used to distinguish members of a set
j	$= \sqrt{-1}$; also used as an integer
$J(f_1)$	a column vector with the m^{th} element $= \cos(2\pi f_1 m \Delta)$
k	an integral-valued index used as a time index; e.g., g_k is the value of g observed at time kT
K	covariance matrix
ℓ_k	$= \ell(X_k \lambda_{k-1})$, the k^{th} value of the likelihood ratio conditioned on the past
$\ell(\cdot)$	the likelihood ratio
L	the loss associated with a false alarm relative to the loss associated with a miss when the loss associated with a correct decision is zero
m	an integral-valued index
M	the number of possible classes
M_c	the memory capacity required of a learning system
$n(t)$	a noise waveform
N	the column vector with samples of $n(t)$ as elements
p_1, p_2	the a priori probability of occurrence of hypotheses 1 and 2 respectively
$p(\cdot)$	a probability density distribution
$P(\cdot)$	a cumulative probability distribution
$\text{Pr}\{\cdot\}$	the probability that the indicated event will occur. Occasionally a subscript is used to indicate the event, such as P_{FA} , which is the probability of a false alarm occurring.
q	an integral-valued index used to distinguish different possible values of the unknown parameter

Q	the number of possible values of the unknown parameter
r_1, r_2	the limits of the range of a
R	a generalized signal-to-noise ratio
$s(t)$	a signal waveform of finite time-bandwidth product (may be lowpass or bandpass)
S	the column vector with samples of $s(t)$ as elements
$S_n(f)$	the noise spectral density
t	the time variable; also used as a subscript on vectors to indicate "transpose"
T	a constant, the period of one observation; the duration of a signal waveform
$T(\cdot)$	a function called a statistic
W	the signal bandwidth if the center frequency of the signal is known; the range of the center frequency if it is unknown
$\{w_i\}$	a set of weights on the taps of a delay line used to synthesize a time-varying linear filter
$x(t)$	the observed, or received, waveform which is to be classified
X	the column vector with samples of $x(t)$ as elements
y_i	$= x_i - x_{i-1}$
Y_i	a binary random variable
z	a dummy variable
$Z_i^{(k)}$	an ordered k -tuple with binary-valued components
α	$= p_2/p_1$, the ratio of a priori probabilities
γ	$= Lp_2/p_1$, the threshold; in one instance γ is used as a dummy time variable
$\gamma(\cdot)$	a dummy function used to obtain performance bounds
Δ	$= 1/(2W)$, the sampling interval
Δ_k	$= \theta_k - \theta_{k-1}$, the k^{th} perturbation in the unknown parameter

ϵ	a small quantity; subscripts are used to distinguish one small quantity from another as necessary
$\zeta(t)$	a complex, lowpass time waveform related to $x(t)$ by the equation $x(t) = \text{Re} \{ \zeta(t) \exp(j\omega_0 t) \}$
$\eta(t)$	a complex, lowpass time waveform related to $n(t)$ in the same way that $\zeta(t)$ is related to $x(t)$
θ	the unknown parameter
λ_k	$= \{X_1, X_2, \dots, X_k\}$ which is used as a shorthand notation to indicate that a probability density is conditioned by the values of the past k observations
$\mu(\cdot)$	a dummy function used to obtain performance bounds
$\nu(\cdot)$	a moment-generating function
$\rho, \rho(\cdot)$	the average risk, a performance measure
τ	a time variable
ϕ	a phase variable, used as the phase of a narrowband signal and as the argument of the complex channel parameter G
$\{\phi_i(\cdot)\}$	a set of independent functions
Φ	the set of all possible values of θ
ω	$= 2\pi f$, the radian frequency variable

OTHER SYMBOLS

$ $	indicates conditioning; e.g., $p(X \theta)$ is the probability density of X conditioned on the value of θ
$\wedge$	indicates the true value of a parameter; e.g., $\hat{\theta}$ is the true value of θ
$\sim$	indicates the real part of the associated symbol; e.g., $\tilde{\eta}(t) = \text{Re} \{ \eta(t) \}$
$\sim$	indicates the imaginary part of the associated symbol; e.g., $\tilde{\eta}(t) = \text{Im} \{ \eta(t) \}$
$\cdot$	indicates "is a member of"; e.g., $\theta \in \Phi$ means θ is a member of the set Φ
$\oplus$	indicates a corruptive noise operation; e.g., $S \oplus N$ indicates that a signal has been corrupted by the addition or multiplication of noise

ACKNOWLEDGMENT

I would like to express my appreciation to Dr. Norman Abramson for his valuable guidance and constant encouragement, to Dr. Thomas Kailath for his careful criticism and helpful suggestions, and to my many colleagues at Stanford and at Sylvania Electronic Systems for many stimulating discussions during the course of this work. The financial support provided by Sylvania is gratefully acknowledged.

I. INTRODUCTION

The purpose of this research has been to apply the techniques of decision theory to the problem of constructing optimal machines which improve their ability to classify patterns by extracting pertinent information from a previously unclassified sequence of observations; such machines learn without a teacher.

In recent years interest in classification problems and in machines to automatically solve these problems has been intensified by the development of a technology in which such problems occur more and more frequently and the development of the analytical and physical tools with which to solve the problems. As a particular example, the advent of the intercontinental ballistic missile has made it mandatory that surveillance systems operate as rapidly and accurately as possible; such systems introduce a variety of classification problems. The development of high-speed large-capacity digital computers has made it possible to perform extremely complex data processing in real time. It is anticipated that both the number of classification problems and the capability of the tools to solve these problems will increase in the next few years.

A. CLASSIFICATION PROBLEMS

In order to be more precise in the meaning of "optimum" and "learning without a teacher," it is necessary to define the classification problem in decision-theory terms: Given an object and a set of classes from which the object may have been drawn, determine the class from which the object was drawn. To get a reasonable solution (by some criterion of reasonableness) one must also be given some knowledge of the losses which will be incurred if an improper determination is made.

In order to solve the problem some set of measurements must be chosen. A particular set of measurements will be called an "observation" and will be represented by a column vector X . (Each element of the vector represents the measurement of a particular parameter, such as "the amplitude of a voltage at time t_0 " or "the color of the object." Thus the specification of a set of measurements may be thought of as the

labeling of the coordinates of an observation space.) For the purposes of this research it is assumed that the observation space is given (that is, it is known which measurements to make), and the problem is to determine a way to process the observations to make a classification decision which is in some sense "optimum."

In order to come to a definition of "optimum," some information must be given regarding the losses associated with misclassification. For this purpose it is assumed that a loss function which provides this information is given. This loss function depends both on the decision to place the object observed in a particular class and the actual class from which the object was drawn. Thus there is a risk associated with each particular decision.

A reasonable definition of the optimum system is that system which minimizes the expected or average risk. Such a system is a realization of a Bayes decision rule, and throughout this report the Bayes system will be considered to be optimum.

When phrased in these terms, classification problems may be characterized in terms of the probability measures induced on the observation space by the different classes of objects. Thus if an object being observed is a member of class 1 the observation X will have a cumulative probability distribution,[†] say $P_1(X)$; if the object belongs to class 2 the observation will have a different cumulative probability distribution, say $P_2(X)$, etc. Three categories of decision problems are possible:

1. The functional forms of the relevant probability measures may be completely known.

[†] Throughout this report it is assumed that the cumulative probability distributions are representable by probability density distributions; e.g., for a scalar observation

$$P(X) = \int_{-\infty}^X p(\alpha) d\alpha ,$$

where the integral is taken in the Lebesgue-Stieltjes sense [Ref. 1] and the class of functions $p(\alpha)$ includes the delta function defined by Middleton [Ref. 2].

2. The functional forms may be completely unknown.
3. The functional forms may be known except for some set of unknown parameters.

Problems which lie in the first category do not involve learning because the solution is defined once the relevant probability measures are known [Refs. 2, 3, 4]. Problems in the second category are commonly referred to as nonparametric. Although there are many important problems in this category (e.g., speech recognition, medical diagnosis, weather prediction) which have been investigated with varying degrees of success, no systematic analytical approach has been developed for such problems. Since it will be assumed that the functional forms of the probability measures are known except for some set of parameters, the techniques developed here will be applicable only to the third, or parametric, class of decision problems.

B. DECISION MACHINES WHICH LEARN

A machine to solve the classification problem must be designed to apply a decision rule to each observation. It seems clear that the decision rule should depend upon how much is known about the problem prior to the time at which the classification decision is to be made. If the problem is a parametric one, it is characterized by a set of probability measures depending on an unknown parameter, say $\{p_i(X|\theta); i = 1, 2, \dots, M\}$, where θ is the parameter. Suppose that this set is known, that an observation is available, and that some a priori knowledge of the parameter [represented by an a priori distribution $p_0(\theta)$] is given. Then a decision rule may be found using standard techniques [Refs. 2, 3, 4].

If, in addition, a sequence of observations, $\{X_1, X_2, \dots, X_k\}$ is available, and if this sequence contains information concerning θ , this information may be extracted and used to modify the decision rule. This may be accomplished by using the sequence of observations (which shall be called a learning sequence and designated by λ_k) to compute a sequence of conditional distributions of θ :

$$p_0(\theta) \rightarrow p(\theta|\lambda_1) \rightarrow p(\theta|\lambda_2) \rightarrow \dots \rightarrow p(\theta|\lambda_k)$$

This sequence of distributions defines a sequence of decision rules, and the resulting machine may be said to "learn."

It is desirable to make a distinction between two modes of learning. A machine which learns with a teacher is provided with two pieces of information: (1) a learning sequence and (2) the correct classification of each member of the sequence. A machine which learns without a teacher is not given the latter information. Thus a machine which learns without a teacher may utilize only that information which is available prior to receiving the first observation or which is contained in the learning sequence. In contradistinction, a machine which learns with a teacher must be externally aided.

There are many problems in which our external means of classification is either poor or nonexistent. If the machine which is built to solve these problems must make repetitive decisions, then sooner or later an observation sequence will become available. If there is any information in one observation concerning other observations, and if we desire a machine which takes advantage of this information, then we require a system which learns without a teacher. (The nature of these problems excludes machines which must be trained.)

There are also problems that require a machine which continues to improve in performance after it has been placed in operation. Included in this class are problems in which the characteristics of the pattern to be recognized are changing with time. A machine could be trained during operation only if the correct classification of each new observation were known, but if this were the case we would not need the machine.

These types of problems provide a compelling motivation for this research which is concentrated on the synthesis of machines which learn without a teacher.

C. RELATED WORK

One of the first engineering problems which led to development of a recognized adaptive system was that of communication through a random channel. In 1956 Price [Ref. 5] and later Price and Green [Ref. 6], using a unique combination of theoretical analysis and engineering

intuition, developed an adaptive receiver called RAKE which effectively reduced the difficulty of communication through random multipath channels by estimating some of the channel properties while receiving signals. Kailath [Refs. 7 and 8] derived an optimum receiver for the same problem and showed that it exhibited adaptive properties. He also pointed out that the RAKE receiver was closely related to the optimum receiver. Proakis and Drouilhet [Ref. 9] simulated two types of binary communication systems using decision-directed feedback to learn the unknown phase of a received signal; they have derived error probabilities which verify that systems of this nature are in some cases superior to nonadaptive systems. Scudder, in 1964 [Ref. 10], derived the optimum learning receiver for the same communication problem; however, in the form he derived, the receiver grows exponentially (see Chapter II). For this reason Scudder proposed and analyzed a decision-directed learning scheme.

In 1961 Glaser [Ref. 11] used a combination of decision-theoretic and intuitive arguments to arrive at an adaptive machine to learn unknown repetitive waveforms in a background of noise. Jackowatz, Shuey, and White [Ref. 12] invented a machine for the same purpose in 1961. Both of these machines learn with a teacher, using decision-directed feedback as the teacher. Hinich [Ref. 13] performed an analysis of the Jackowatz machine in 1962, and later [Ref. 14] modified the mathematical model to obtain a more precise analysis.

The work which is most closely related to the research presented here was initiated by Braverman in 1961 [Ref. 15]. Braverman examined the problem in which a previously classified learning sequence is available (learning with a teacher), and established the fact that the solution which uses all relevant observations to condition the a posteriori probabilities, achieves the minimum average risk. He also established the convergent properties of this solution. This work was extended by Abramson and Braverman [Ref. 16] and applied to the problem of learning the vector mean of a random vector which was normally distributed. In 1963 Keehn [Ref. 17] solved the more general learning problem in which the random vector is normally distributed with both unknown mean and

covariance matrix. At the same time Spragins [Ref. 18] generalized the approach of Abramson and Braverman in a different direction to obtain the solution to the general parametric learning (with a teacher) problem in which a fixed-size (nontrivial) sufficient statistic exists.

While work on the learning with a teacher problem was continuing, Daly [Refs. 19 and 20] in 1961 used a decision-theory approach (the "Bayes" approach used by Braverman) to attack the learning without a teacher problem. Daly solved the one-dimensional binary detection problem and established the convergence of the solution; however, he also demonstrated that his solution required a system which grew exponentially with the number of learning observations. Both Daly and Scudder turned their attention to systems which, like the majority of those proposed to solve the learning problem without a teacher, use decision feedback as a teacher to aid in the learning process. A more complete explanation of the exponentially growing system will be found in Chapter II.

In this investigation we have taken the so-called Bayes approach (explained in Chapter II) which was used by other investigators [Refs. 10, 15-20], and we have concentrated on the learning without a teacher problem. One of the most important results is the fact that in many problems this approach does lead to systems of fixed size. In problems in which the system size must grow, a change in the formulation of the problem will result in fixed-size systems. This change requires that we approximate the space of the unknown parameter; however the performance of the resulting fixed-size system is in an engineering sense equivalent to the growing system.

D. ORGANIZATION, APPROACH, AND SIGNIFICANT RESULTS

In the first portion of this report (Chapters II and III), the equations describing the learning system are derived and then applied to signal-detection problems in which an important signal parameter is unknown but fixed. The performance of such a system is discussed. In the second part, which consists of Chapters IV, V, and VI, the equations are applied to problems in which the important unknown parameter is time varying. The stability, convergence, and realizability of the general learning system are discussed.

The investigation of the learning problem is initiated by defining a suitably general, repetitive binary decision problem depending upon an unknown parameter which is fixed. This parameter is treated as a random variable, and an a priori probability distribution is chosen to describe the initial state of knowledge of the parameter. As more and more observations are received, more and more information concerning the parameter is obtained; thus the observations "condition" the probability distribution of the parameter. By developing a recursive expression which describes this unfolding sequence of conditional probability distributions, a mathematical description of the learning process is obtained; and by utilizing this recursive expression, a learning system is synthesized.

This technique is applied to two examples in order to illustrate the types of problems that are readily solved. The first example involves the detection of a signal of known waveform but unknown amplitude embedded in noise. It has been chosen to illustrate the technique as simply as possible. The second example involves the detection of a narrowband signal of unknown frequency embedded in noise. It has been chosen as an example of a problem which frequently occurs in the electronic-countermeasures and reconnaissance fields, which is readily solved by the proposed technique, but which has not been attacked successfully by any other means [Ref. 21].

The investigation is continued by extending the development to (1) the "learning" problem in which the a priori probabilities of occurrence of the alternative hypotheses are unknown, and (2) the repetitive multiple-hypothesis decision problem.

In the third chapter techniques for the evaluation of system performance are discussed briefly. An example is presented of the performance bounds of a system that detects the presence of a narrowband signal of unknown frequency in bandlimited white gaussian noise. Thus this latter example, which is used in both Chapters II and III, may be used to present a sort of overview of the major contribution of this research to the reader familiar with the so-called Bayes approach to the decision problem.

Succeeding chapters are extensions of the solution and developments of the properties of the resulting system.

In the fourth chapter the technique is extended to include problems in which the unknown parameter is randomly varying in time. Depending upon the model of time variations, the resulting systems are simple modifications of the learning systems for fixed parameters. The synthesis technique is applied to examples of communications, radar, and electronic reconnaissance problems.

In the fifth chapter the size of the optimal learning system is defined in terms of the number of elements required to construct the system. It is shown that in many cases the optimal systems are of finite size. In the cases where the optimum systems grow as the learning sequence lengthens, it is shown that a suboptimum system can be constructed from a finite number of elements. In the sixth chapter it is shown that the finite suboptimum system has a performance which is not measurably different from the optimum system. Thus from an engineering standpoint the optimum learning system may always be realized from a finite number of elements.

Other properties of learning systems are presented in the sixth chapter. The systems are stable and converge in performance so that as the learning sequence lengthens the performance of the learning system is equivalent to the performance of a system which is given a priori knowledge of the unknown parameter.

II. DEVELOPMENT OF THE LEARNING SYSTEM

As was pointed out in Chapter I, there are many repetitive decision problems in which an important parameter is unknown and in which external aid is not available from which to obtain a representative set of properly classified "learning" observations. Such problems require systems which learn without a teacher, and it is the purpose of this chapter to develop and explain a general technique for the synthesis of such systems. The technique developed is based on the assumption that the unknown parameters are either time invariant, or so slowly time varying that they may be treated as being fixed. The more difficult time-varying unknown-parameter problem is investigated in Chapter IV.

A. THE LEARNING PROBLEM MODEL (BINARY DECISIONS)

In order to explain the techniques involved in the synthesis of learning systems, we first consider the binary decision problem. We shall phrase the problem in terms of detection of a signal which depends upon a set of unknown parameters; however, the result will be easily generalized.

Assume that we are given an observation representable by the column vector X and a learning sequence $\lambda_{k-1} = \{X_1, X_2, \dots, X_{k-1}\}$, consisting of the first $(k-1)$ such vectors. Each observation contains a signal corrupted by noise, or it contains noise alone, and we desire to synthesize a system to decide whether or not the k^{th} observation X_k contains a signal. We assume that our system may make mistakes, and that each mistake costs something which may be expressed in terms of a function which depends on the actual situation as well as the decision. We ask for a system which will minimize the average risk associated with each decision; i.e., we require a system which is optimum in the Bayes sense at every decision instant. The system which performs this minimization computes the likelihood ratio and compares it to some threshold. To be more precise, we let

H_1 = the hypothesis that $X_k = S(\theta) \oplus N$

H_2 = the hypothesis that $X_k = N$

where

$S(\theta)$ = the signal vector (unknown parameters)

θ = unknown signal parameters

N = the noise vector

$\oplus$ = the corrupting operation (addition, multiplication, etc.)

Then, if the signal parameters were known, the optimum system would compute the ratio of conditional probabilities, or likelihood ratio [see Ref. 4]:

$$\ell(X_k|\theta) = \frac{p(X_k|H_1, \theta)}{p(X_k|H_2)} \quad (2.1)$$

and compare it to a threshold depending upon the relative loss associated with the two types of errors (false alarm and miss) and the a priori probability of occurrence of the two hypotheses.

If the signal were random with known distribution $p(\theta)$, the optimum system would compute an average likelihood ratio [see Ref. 4]:

$$\ell(X_k) = \int \ell(X_k|\theta) p(\theta) d\theta \quad (2.2)$$

In the problem at hand, when we wish to take advantage of all prior information, we may easily show that the optimum system computes a conditional likelihood ratio [Ref. 15, pp. 12-16], that is, a ratio of probabilities conditioned on the past:

$$\ell(X_k|\lambda_{k-1}) = \frac{p(X_k|\lambda_{k-1}, H_1)}{p(X_k|\lambda_{k-1}, H_2)} \quad (2.3)$$

This may be rewritten in a more useful form as a conditional expectation:

$$\ell(x_k | \lambda_{k-1}) = \int \ell(x_k | \theta) p(\theta | \lambda_{k-1}) d\theta \quad (2.4)$$

We have assumed that θ is the only unknown parameter; hence, if we were given θ we could learn nothing from λ_{k-1} ; thus $\ell(x_k | \theta, \lambda_{k-1}) = \ell(x_k | \theta)$.[†] The synthesis of a system which will compute this latter function is a standard problem of detection theory, and solutions are usually known. The problem of interest involves the synthesis of a system to compute $p(\theta | \lambda_{k-1})$.

B. AN EXPONENTIALLY GROWING SOLUTION

In order to understand the difficulty which arises when we attempt to synthesize a system to compute $p(\theta | \lambda_{k-1})$, we may take the following approach (suggested by Daly [Refs. 19, 20] and Scudder [Ref. 10]). Suppose that we knew which members of the sequence $\{X_1, X_2, \dots, X_{k-1}\}$ contained a signal. Then we could use these members in a machine which learned with a teacher (Refs. 14, 15, and 16 tell us how to construct such machines). If we lack knowledge to learn with a teacher, we may still build 2^{k-1} machines which learn with a teacher, partition the sequence into the 2^{k-1} possible ways in which the $(k-1)$ observations might be classified, and feed one partition into each machine. Each partition will have a known probability of occurrence; thus if we weight the output of the 2^{k-1} learning machines by the appropriate probabilities of occurrence and sum, we will have solved the problem. Clearly, the resulting system will grow exponentially as we add more learning observations.

[†]This amounts to an assumption of conditional independence which may be written

$$p(X_1, X_2, \dots, X_k | \epsilon, H_1) = p(X_1 | \epsilon, H_1) p(X_2 | \epsilon, H_1) \dots p(X_k | \epsilon, H_1)$$

To be more precise, we define a binary random vector $Z_i^{(k)}$ such that it has k components. Each component has the value 1 or 0 depending respectively upon whether the signal is or is not present in the observation which that component represents. [The sequence $\langle \text{present, not present, present} \rangle$ is represented by the vector $\langle 101 \rangle = Z^{(3)}$.] $Z_i^{(k)}$ is an ordered k -tuple with binary-valued components, and there are 2^k possible $Z_i^{(k)}$. These may be ordered by letting $Z_i^{(k)}$ equal the binary expansion of i as i varies from 0 to $2^k - 1$. By conditioning the distribution $p(\theta | \lambda_{k-1})$ on the random vector $Z_i^{(k)}$, and averaging over all i , we obtain

$$p(\theta | \lambda_{k-1}) = \sum_{i=0}^{2^{k-1}-1} p(\theta | \lambda_{k-1}, Z_i^{(k-1)}) p(Z_i^{(k-1)} | \lambda_{k-1}) \quad (2.5)$$

Thus (2.4) may be rewritten as

$$\ell(x_k | \lambda_{k-1}) = \sum_{i=0}^{2^{k-1}-1} p(Z_i^{(k-1)} | \lambda_{k-1}) \int \ell(x_k | \theta) p(\theta | \lambda_{k-1}, Z_i^{(k-1)}) d\theta \quad (2.6)$$

Both of the conditional distributions in (2.6) may be expanded in terms of known functions; e.g.,

$$p(Z_i^{(k-1)} | \lambda_{k-1}) = \frac{p(\lambda_{k-1} | Z_i^{(k-1)}) p(Z_i^{(k-1)})}{\sum_{i=0}^{2^{k-1}-1} p(\lambda_{k-1} | Z_i^{(k-1)}) p(Z_i^{(k-1)})} \quad (2.7a)$$

$$p(\lambda_{k-1} | Z_i^{(k-1)}) = \int p(\lambda_{k-1} | Z_i^{(k-1)}, \theta) p_o(\theta) d\theta \quad (2.7b)$$

$$p(\lambda_{k-1} | Z_i^{(k-1)}, \theta) = \prod_{j=1}^{k-1} p(X_j | Z_i^{(k-1)}, \theta) \quad (2.7c)$$

$$p(X_j | Z_i^{(k-1)}, \theta) = \begin{cases} p(X_j | H_1, \theta) & \text{if } j^{\text{th}} \text{ component of } Z_i^{(k-1)} = 1 \\ p(X_j | H_2) & \text{if } j^{\text{th}} \text{ component of } Z_i^{(k-1)} = 0 \end{cases} \quad (2.7d)$$

Thus from (2.6) a system may be synthesized. Unfortunately, the system will grow exponentially as the learning sequence lengthens. That is, 2^k computations are required for the optimum utilization of k learning observations. As the length of the learning sequence increases, the system grows in size very rapidly, and for this reason it does not seem practical for large values of k . When we are interested only in small values of k , however, this type of system may be quite practical, and may even result in a less complex system than the one which we shall describe in the next section.

The conclusion that optimal machines which learn without a teacher are impractical for large k might seem to follow from the above argument. In fact, however, this is not the case as will be demonstrated in the next paragraph.

C. A RECURSIVE SOLUTION

In order to obtain a system of fixed size, we return our attention to $p(\theta | \lambda_{k-1})$ and proceed as follows.

$$p(\theta | \lambda_{k-1}) = p(\theta | X_1, X_2, \dots, X_{k-1}) \quad (2.8a)$$

or by Bayes' law,

$$\begin{aligned} p(\theta | \lambda_{k-1}) &= \frac{p(X_{k-1} | \theta, X_1, \dots, X_{k-2}) p(\theta | X_1, \dots, X_{k-2})}{p(X_{k-1} | X_1, \dots, X_{k-2})} \\ &= \frac{p(X_{k-1} | \theta, \lambda_{k-2}) p(\theta | \lambda_{k-2})}{p(X_{k-1} | \lambda_{k-2})} \end{aligned} \quad (2.8b)$$

Consider the conditional density $p(X_{k-1} | \theta, \lambda_{k-2})$. There are two possibilities for X_{k-1} : H_1 may be true or H_2 may be true. Thus $p(X_{k-1} | \theta, \lambda_{k-2})$ may be written as a mixture:

$$p(X_{k-1} | \theta, \lambda_{k-2}) = p(H_1) p(X_{k-1} | H_1, \theta, \lambda_{k-2}) + p(H_2) p(X_{k-1} | H_2, \theta, \lambda_{k-2}) \quad (2.9)$$

In Eq. (2.9) we have assumed that $p(H_1)$ and $p(H_2)$ are known a priori. In many interesting problems this is not the case. The problem where $p(H_1)$ and $p(H_2)$ are not known a priori is treated later in this chapter. If H_1 is true and θ is known, then X_{k-1} does not depend on λ_{k-2} . The noise is independent of the signal; therefore, if H_2 is true, X_{k-1} does not depend on either θ or λ_{k-2} ; thus

$$p(X_{k-1} | \theta, \lambda_{k-2}) = p(H_1) p(X_{k-1} | \theta, H_1) + p(X_{k-1} | H_2) p(H_2) \quad (2.10)$$

By similar reasoning we may write

$$p(X_{k-1} | \theta, \lambda_{k-2}) = p(H_1) p(X_{k-1} | H_1, \lambda_{k-2}) + p(H_2) p(X_{k-1} | H_2) \quad (2.11)$$

Finally, by factoring and rewriting (2.8b), (2.10), and (2.11) we have

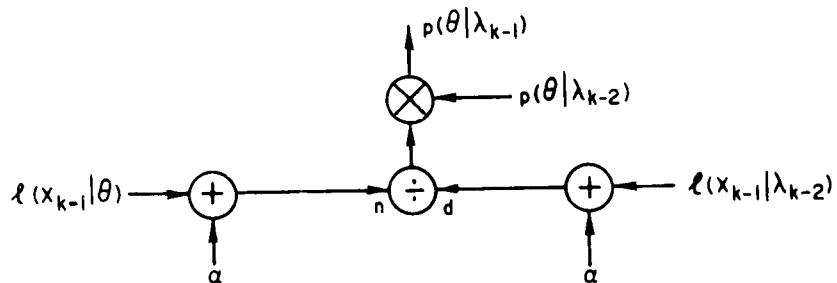
$$p(\theta | \lambda_{k-1}) = \left[\frac{p(X_{k-1} | \theta) + \alpha}{p(X_{k-1} | \lambda_{k-2}) + \alpha} \right] p(\theta | \lambda_{k-2}) \quad (2.12)$$

where $\alpha = p(H_2)/p(H_1)$. The importance of Eq. (2.12) lies in the fact that it has a recursive form. This fact will allow synthesis of a system in delay-feedback form. As discussed in Chapter V, the system may be realized if the number of usefully distinguishable possible values of θ is finite.

If we review the computations required of this system which learns without a teacher, we find that we are in a position to synthesize the system. The computations required are:

1. Compute $\ell(X_k|\theta)$ for each possible θ
2. Compute $p(\theta|\lambda_{k-1})$ for each possible θ
3. Weight (1) by (2) and sum over all θ .

The third computation will result in $\ell(X_k|\lambda_{k-1})$. We have assumed that we know how to compute $\ell(X_k|\theta)$. Suppose that somehow we could obtain $\ell(X_{k-1}|\lambda_{k-2})$ and $p(\theta|\lambda_{k-2})$, then the system of Fig. 1 would provide the desired $p(\theta|\lambda_{k-1})$.

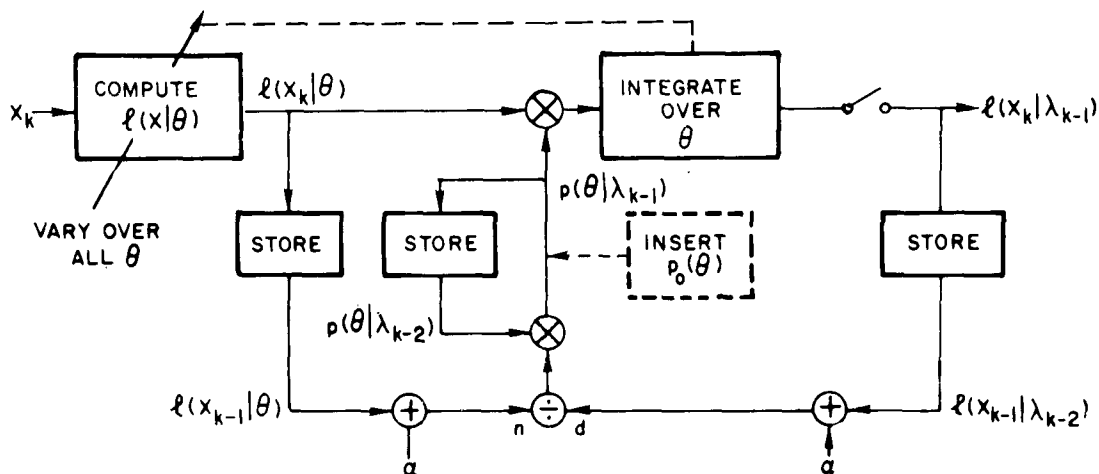


33374

FIG. 1. A SYSTEM TO COMPUTE $p(\theta|\lambda_{k-1})$.

In Fig. 1 and throughout this report, the symbol  has been used to indicate a zero-memory device which has as an output the ratio of the two inputs. The input marked "n" is the numerator, that marked "d" is the denominator. The multipliers  and adders  are also zero-memory devices.

If we simply store $p(\theta|\lambda_{k-1})$, we will have it available for computation of $p(\theta|\lambda_k)$ when the next vector X_{k+1} is received. Similarly, if we store $\ell(X_k|\theta)$, it will be available when X_{k+1} is received. The system shown in Fig. 2 is one form of the required learning system.



33373

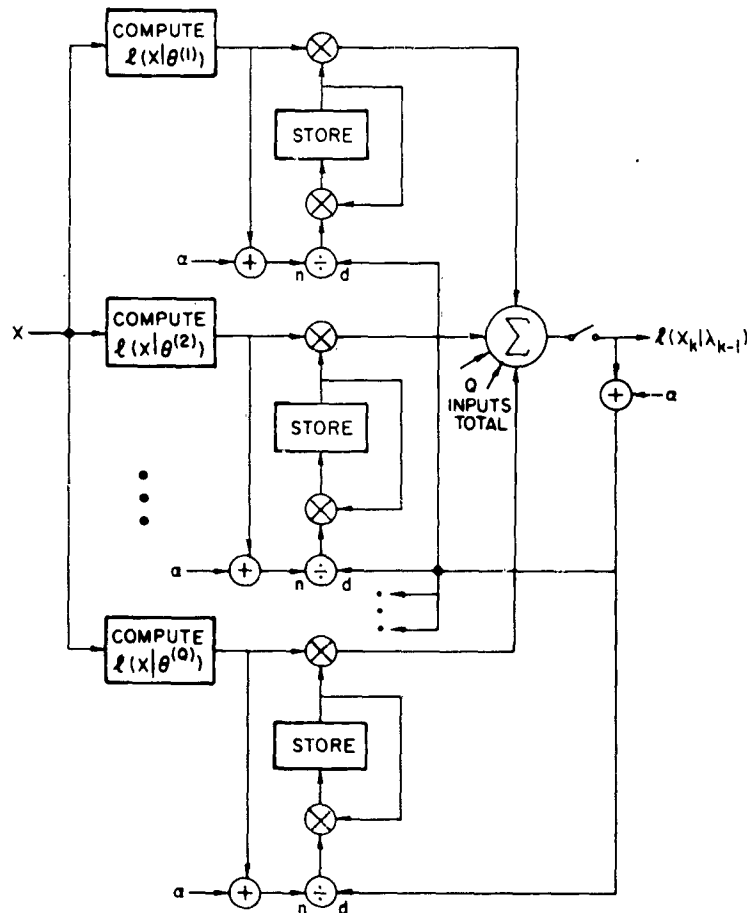
FIG. 2. A BINARY DETECTOR WHICH LEARNS WITHOUT A TEACHER (SEQUENTIAL FORM).

There are several facts concerning this system which, although self-evident, should be considered. First, the computation of $l(x|\theta)$ and $p(\theta|\lambda_{k-1})$ must be made for every possible value of θ . The integrator must be synchronized with the sequential variation of θ . Second, when the machine is started an initial distribution of θ , or $p_0(\theta)$, must be inserted. This distribution may be uniform over θ , or it may have any convenient form consistent with our a priori knowledge of θ .

The fact that the computation of $l(x|\theta)$ and $p(\theta|\lambda_{k-1})$ must be made for every possible value of θ poses a difficult problem. If θ varies in a continuous space, there will be an uncountable infinity of possible values, and the various components of Fig. 2 will not be realizable exactly. We shall circumvent this problem by assuming that the space of θ can be quantized, so that the system need compute $l(x|\theta)$ for only a finite number of values of θ . Later, in Chapter V, we shall demonstrate that a quantized space may always be chosen so that the performance of a system based on this space will be arbitrarily close to the performance of the theoretical system.

The assumption that the space of θ has a finite number of points allows us to represent the system in an alternative form as illustrated in Fig. 3. In this form the system computes $l(x|\theta)$ and $p(\theta|\lambda_{k-1})$ and takes the product simultaneously for all values of θ . The products thus

formed are summed, and the result is $\ell(X_k | \lambda_{k-1})$. Both the sequential and the parallel forms of the system will be used in the various examples in this and following chapters.



33365

FIG. 3. A BINARY DETECTOR WHICH LEARNS WITHOUT A TEACHER (PARALLEL FORM).

D. EXAMPLES

The system which has just been derived may be applied directly to a wide variety of signal-detection problems. This application requires only that we determine an explicit expression for $\ell(x|\theta)$ and synthesize a system to compute the expression. The following examples demonstrate this procedure.

1. Detection of a Signal of Unknown Amplitude

In this example we shall consider the problem of detecting a signal of known waveform but unknown amplitude embedded in additive noise. Such a problem might arise if we were to use for a communication channel a medium which faded so slowly that the attenuation could be considered constant. (For a more realistic consideration of the fading-channel communication problem, and an application of this example, see example 1 of Chapter IV.)

We assume that the signal to be detected may be written as the product of an unknown scalar and a known bandlimited waveform of duration T .

$$s(t) = cb(t)$$

where $b(t)$ = known waveform of bandwidth W , duration T

c = unknown scalar

The signal is embedded in a background of additive, gaussian noise of zero mean and covariance matrix K .

In our hypothetical problem we are given a received waveform $x(t)$ (perhaps an i-f amplifier voltage) which starts at time zero and continues to the present. For simplicity we assume that the signal is of duration T and may only start at instants separated from a known synchronization instant by integral multiples of T . The signal is transmitted at intervals chosen at random, and our problem is to look at the received waveform for a duration T $\{x(t); (k-1)T \leq t \leq kT\}$ and decide whether or not the signal is present.

In order to easily manipulate the appropriate variables we shall take advantage of the representation of continuous bandlimited waveforms by vectors of the sample values of the waveform. We shall denote

$$S = \begin{bmatrix} s(0) \\ s\left(\frac{1}{2W}\right) \\ \vdots \\ s\left(T - \frac{1}{2W}\right) \end{bmatrix} \quad B = \begin{bmatrix} b(0) \\ b\left(\frac{1}{2W}\right) \\ \vdots \\ b\left(T - \frac{1}{2W}\right) \end{bmatrix} \quad N = \begin{bmatrix} n(0) \\ n\left(\frac{1}{2W}\right) \\ \vdots \\ n\left(T - \frac{1}{2W}\right) \end{bmatrix}$$

where $n(t)$ is the noise waveform.

We divide the received waveform into "observations" of duration T , and denote these by the indexed vectors X_k :

$$X_k = \begin{bmatrix} x((k-1)T) \\ \vdots \\ x(kT - \frac{1}{2W}) \end{bmatrix}$$

We define two hypotheses which apply to each observation:

$$H_1 = \text{the hypothesis that } X_k = cB + N_k$$

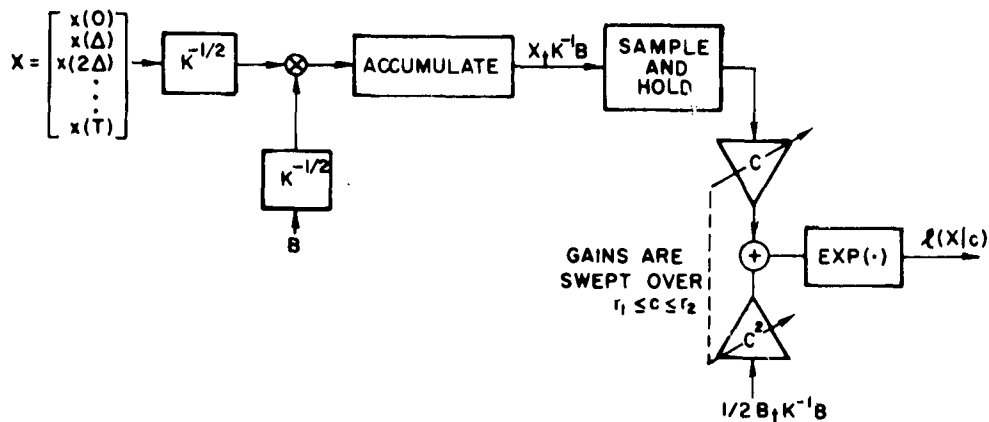
$$H_2 = \text{the hypothesis that } X_k = N_k$$

The optimum learning system must compute the likelihood ratio, $\ell(X|c)$, which is given by

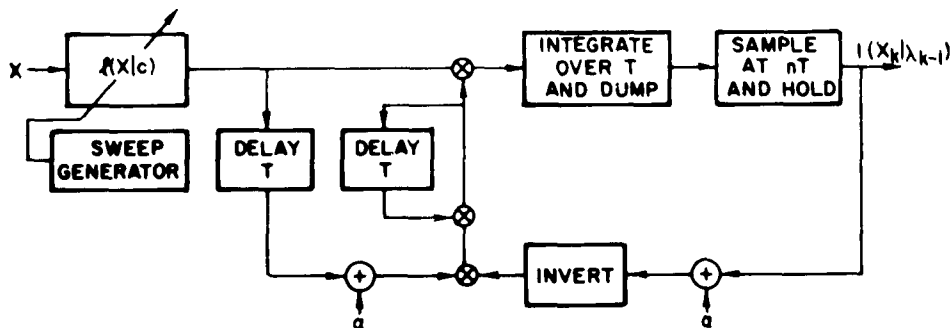
$$\ell(X|c) = \frac{p(X|c, H_1)}{p(X|c, H_2)} = \exp \left(-\frac{1}{2} c^2 B_t K^{-1} B + c X_t K^{-1} B \right)$$

In order to vary the computation over all c , we restrict c to some range, say $r_1 \leq c \leq r_2$. To easily construct the system we make c a function of time, and integrate over time instead of c ; that is, we sweep c linearly from r_1 to r_2 . If we make the sweep period T the same as the observation interval, synchronization will be much more easily achieved. The resulting system is shown in Fig. 4.

In this system the input vector X is transformed by the matrix operator $K^{-1/2}$ to yield the vector $K^{-1/2}X$. This vector is multiplied, term by term, by the vector $K^{-1/2}B$ and the terms summed (accumulated for T sec) to provide $X_t K^{-1}B$. This product is a scalar and is the value of the accumulator sampled at the appropriate instant. This sample is held for T sec while the gains c are swept through the range. During this period the contents of the accumulator are dumped and the



a. The sweeping likelihood computer



b. The detection system

33364

FIG. 4. A LEARNING SYSTEM FOR DETECTION OF SIGNALS OF UNKNOWN AMPLITUDE.

product involving the next observation is accumulated. Thus once every T sec the parameter c is swept through its range (r_1 to r_2) and the output $\lambda(X|c)$ is swept through the range of c .

2. Detection of a Narrowband Signal of Unknown Frequency

A problem which arises often in the different phases of electronic-countermeasures, reconnaissance, and communications fields is the detection of a narrowband signal of unknown frequency f , random amplitude, and

random phase. The problem of detecting such a signal when only the current observation is used has been discussed by Helstrom [Ref. 22][†] and Wainstein and Zubakov [Ref. 23] among others. These authors suggest the use of a receiver which uses a bank of narrowband filters centered at each possible frequency. The filter output which is maximum is compared to a threshold to determine whether a signal is present or not. The performance of such a receiver (called a "maximum likelihood" receiver by Helstrom) is evaluated by Wainstein and Zubakov. Such a receiver is shown to have performance which is nearly as good as the performance of the Bayes or average-likelihood receiver [Ref. 23] without learning. In many cases, however, this performance is not adequate (see Fig. 10 of Chapter III) and it is desirable to take advantage of the fact that the signal is recurring at the same frequency; that is, it is desirable to apply the techniques of learning to the problem.

If we did not desire to use more than k past observations to learn f , we could use 2^k receivers constructed to learn the frequency with a teacher as explained in Sec. B. (See also Refs. 19, 20, and 24.) However for even moderate k such a receiver would be impractical.

In the following paragraphs we shall derive the optimum learning receiver for this problem. We shall see that it consists essentially of a bank of periodogram calculators (which are approximately narrowband filters) whose outputs are the inputs to a bank of antilog devices. The antilog device outputs are weighted by the learned probability distribution of frequency and summed. The sum is the desired likelihood ratio. Mathematically, we proceed as follows.

Assume that the signal $s(t)$ may be represented by a sample function of a narrowband gaussian random process over the interval T when the signal is present. The sample functions are independent from one interval to the next, and the occurrence of a signal in one interval is independent of its occurrence in other intervals. We also assume that the signal can start only at times which are separated from a known

[†] For other discussions of this problem see Refs. 21 and 24.

synchronization instant by integral multiples of T . Thus over any interval T the signal may be described by Eq. (2.13).[†]

$$s(t) = a \cos (\omega t + \phi) \quad (2.13)$$

where a is a random variable with Rayleigh distribution:

$$p(a) = \begin{cases} \frac{a}{A^2} \exp \left(-\frac{a^2}{2A^2} \right) & a \geq 0 \\ 0 & a < 0 \end{cases} \quad (2.14)$$

ϕ is a uniformly distributed random variable

$$p(\phi) = \begin{cases} \frac{1}{2\pi} & 0 \leq \phi \leq 2\pi \\ 0 & \text{elsewhere} \end{cases} \quad (2.15)$$

$f = \omega/2\pi$ is unknown, except that it must be one of a discrete set of frequencies $\{f_1, f_2, \dots, f_Q\}$

The assumption of synchronization may be relaxed, and the synchronization time treated as an unknown parameter. The problem becomes much more complex, and would not serve as a good illustration at this point. An alternative technique when synchronization is unknown is to choose the interval T to be very short compared to the signal duration, and to take into account the resulting signal dependence from interval to interval. This latter approach may be accomplished by treating the probability $p(H_1)$ as a time-varying random parameter, thus combining the techniques of this and the next chapter.

[†] See Refs. 25, 26 for a description of the properties of narrowband gaussian random processes. Equation (2.13) merely expresses the fact that such a process may be described in terms of two independent random variables, the amplitude and the phase.

The noise is assumed to be additive and normally distributed with covariance matrix K .

Because the problem is to observe the received waveform over intervals of duration T and to make repetitive decisions at the end of each interval, we define two hypotheses which apply to each observation:

H_1 = the hypothesis that $x(t) = s(t) + n(t)$

H_2 = the hypothesis that $x(t) = n(t)$

where $x(t)$ = received waveform

$n(t)$ = noise

The optimum learning system must compute the conditional likelihood ratio:

$$\ell(X_k | \lambda_{k-1}) = \sum_{i=1}^Q \ell(X_k | f_i) P(f_i | \lambda_{k-1}) \quad (2.16)$$

where X_k is the k^{th} 2TW-dimensional column vector of samples of $x(t)$ sampled at the interval $\Delta = 1/(2W)$, and W is the bandwidth within which the frequency must fall.

To synthesize a system to solve this problem, we must express $\ell(X | f_i)$ and $P(f_i | \lambda_{k-1})$ explicitly. This may be done as follows.

First we express $\ell(X_k | a, \phi, f_i)$ explicitly and then average over a and ϕ . To this end we may use $p(a)$ and $p(\phi)$ as defined in Eqs. (2.14) and (2.15) since the current (k^{th}) values of a and ϕ are independent of each other and of previous and future values of a and ϕ . Thus

$$\ell(X_k | f_i) = \int_0^{\infty} p(a) \left[\int_0^{2\pi} p(\phi) \ell(X_k | a, \phi, f_i) d\phi \right] da \quad (2.17)$$

Because of the normality of the noise, this integral may be carried out to yield

$$\ell(X_K | f_1) = \frac{C_1}{A^2} \exp \left[\frac{C_1}{2} |X_{K_t} K^{-1} E(f_1)|^2 \right] \quad (2.18)$$

where

$$C_1 = \frac{A^2}{A^2 R_1^2 + 1}$$

$$E(f_1) = \begin{bmatrix} 1 \\ \exp(j2\pi f_1 \Delta) \\ \vdots \\ \exp(j2\pi f_1 T) \end{bmatrix}; \quad J(f_1) = \begin{bmatrix} 1 \\ \cos(2\pi f_1 \Delta) \\ \vdots \\ \cos(2\pi f_1 T) \end{bmatrix}$$

$$R_1 = J_t(f_1) K^{-1} J(f_1)$$

If the noise is stationary, $K^{-1}E(f_1)$ represents the sampled-data form of the output of a linear filter with system function the reciprocal of the noise spectral density when $\exp(j\omega_1 t)$ is the input. Thus the effect of assuming that the noise is not white may be taken into account by the introduction of a multiplicative constant (depending on f_1) in the exponent. Let $S_N(f)$ be the noise spectral density, then

$$\ell(X_K | f_1) = \frac{C_1}{A^2} \exp \left[\frac{C_1}{2} D_1 |X_t E(f_1)|^2 \right] \quad (2.19)$$

where $D_1 = 1/(S_N(f_1))$. The quantity $|X_t E(f_1)|$ may be recognized as being proportional to the sampled-data form of the periodogram of $x(t)$, since

$$|X_t E(f_i)| = \left| \sum_{m=1}^{2TW} x(m\Delta) \exp(j\omega_i m\Delta) \right|^2 \doteq 4W^2 \left| \int_0^T x(t) \exp(j\omega_i t) dt \right|^2 \quad (2.20)$$

and the periodogram (at $f = f_i$) is defined as

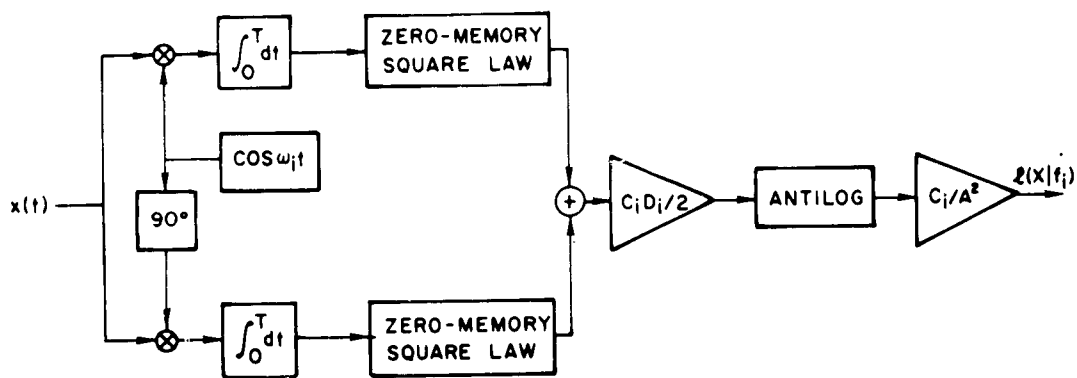
$$\text{Per}_i [x(t)] = \frac{1}{T} \left| \int_0^T x(t) \exp(j\omega_i t) dt \right|^2 \quad (2.21)$$

Hence the system is required to compute the periodogram of $x(t)$ at each of the frequencies $\{f_i; i = 1, 2, \dots, Q\}$, to weight each of these computations by $C_i D_i / 2$, to take the antilog of the result, and to weight the antilog by C_i / A^2 . This operation may be performed sequentially by a single circuit or in parallel by a bank of Q circuits. (If it is performed sequentially and if the frequencies are taken to be a set within the band W separated by $1/2T$ cps, the circuit required may be identified as a form of time-compressive sweeping receiver which sweeps the band W in time T with resolution $1/T$ [Refs. 23, 24]. The construction of such a receiver is quite possible; however it may be somewhat confusing to introduce the concept at this point and therefore we shall utilize the parallel form of receiver.)

Since we have found the form of $\mathcal{L}(X|f_i)$, the problem is solved in the parallel form by inserting the $\mathcal{L}(X|f_i)$ computer, defined by Eq. (2.18) and illustrated in Fig. 5, into the appropriate box in Fig. 3 (identifying θ_i of Fig. 3 with f_i of Fig. 5). The result is the system of Fig. 6.

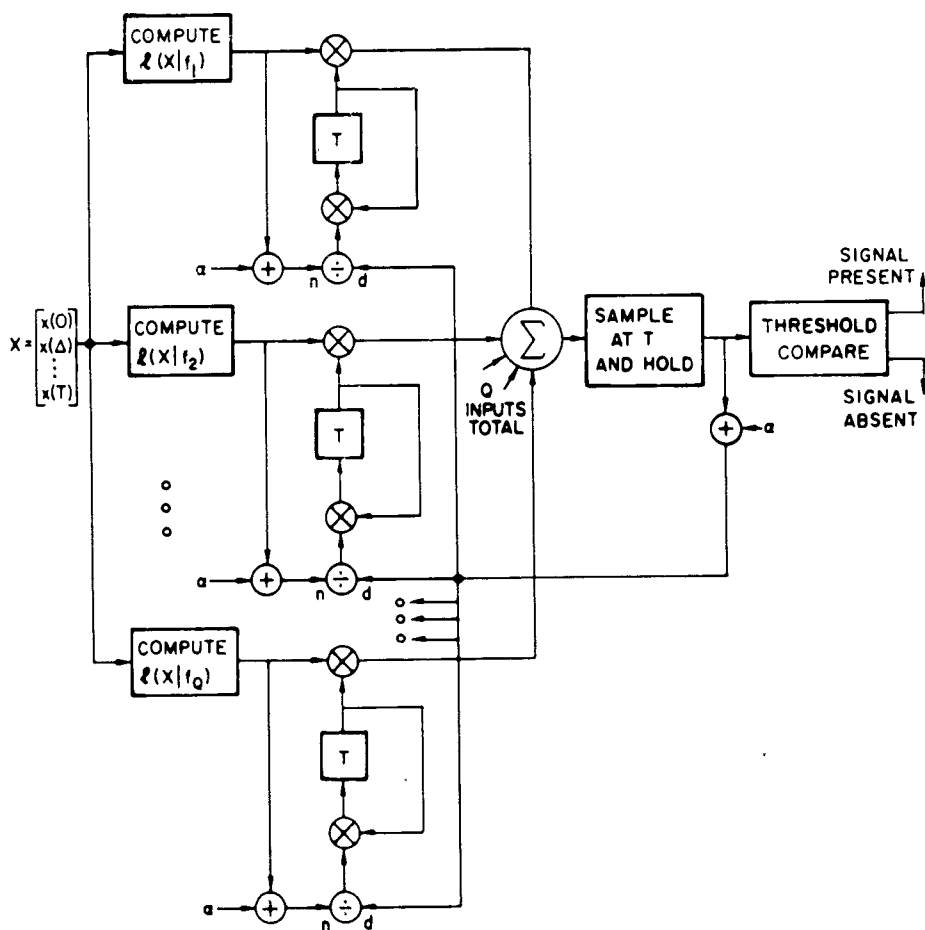
E. LEARNING THE A PRIORI PROBABILITIES

In the model originally proposed it was assumed that the a priori probabilities $p(H_1)$ and $p(H_2)$ were known but that some other parameter was unknown. In many problems only the a priori probabilities are unknown, and in other problems both the a priori probabilities and other parameters are unknown.



33375

FIG. 5. A COMPUTER FOR $l(x|f_i)$.



33363

FIG. 6. A LEARNING RECEIVER.

The solution to such problems cannot be obtained by treating $p(H_1)$ and $p(H_2)$ as signal parameters because they do not appear in the equations in the same manner. Since it is the purpose of this section to outline the proper solution, we shall begin by assuming that $p(H_1) = p_1$ and $p(H_2) = p_2$ are the only unknown parameters. If they were known, the optimum system could be realized by computing $\ell(X_k)$ and comparing it to a threshold Lp_2/p_1 as previously noted. However, another optimum system is one which computes $(p_1/p_2)\ell(X_k)$ and compares it to L . By utilizing this latter system we may show that when $p(H_1)$ and $p(H_2)$ are unknown but a sequence of learning observations λ_{k-1} is available, the optimum system computes the conditional expectation of $(p_1/p_2)\ell(X_k)$, defined in (2.22), and compares it to the threshold L [Ref. 14]. The conditional expectation is

$$L(X_k | \lambda_{k-1}) = \int \ell(X_k) \frac{p_1}{1 - p_1} p(p_1 | \lambda_{k-1}) dp_1 \quad (2.22)$$

where we have taken advantage of the fact that $p_2 = 1 - p_1$. Now following the procedure which led to Eq. (2.12) we have by Bayes' law

$$p(p_1 | \lambda_{k-1}) = \frac{p(X_{k-1} | p_1, \lambda_{k-2}) p(p_1 | \lambda_{k-2})}{\int p(X_{k-1} | p_1, \lambda_{k-2}) p(p_1 | \lambda_{k-2}) dp_1} \quad (2.23)$$

Since p_1 is the only unknown variable, we may write

$$\begin{aligned} p(X_{k-1} | p_1, \lambda_{k-2}) &= p(X_{k-1} | H_1, \lambda_{k-2}, p_1) p(H_1 | p_1, \lambda_{k-2}) \\ &+ p(X_{k-1} | H_2, \lambda_{k-2}, p_1) p(H_2 | p_1, \lambda_{k-2}) \end{aligned} \quad (2.24)$$

But when we know that X_{k-1} comes from the class of observations which contain signal, we know the probability density function of X_{k-1} ; hence

$$p(X_{k-1} | H_1, \lambda_{k-2}, p_1) = p(X_{k-1} | H_1) \quad (2.25a)$$

Similarly,

$$p(X_{k-1}|H_2, \lambda_{k-2}, p_2) = p(X_{k-2}|H_2) \quad (2.25b)$$

that is, the observations are conditionally independent of the past and of the value of p_1 when either H_1 or H_2 is given. When the value of p_1 is known, $p(H_1)$ and $p(H_2)$ are known, so that

$$p(H_1|p_1, \lambda_{k-2}) = p_1 \quad (2.26a)$$

$$p(H_2|p_1, \lambda_{k-2}) = 1 - p_1 \quad (2.26b)$$

From Eqs. (2.24), (2.25a), and (2.25b) we may write $p(p_1|\lambda_{k-1})$ in terms of the likelihood ratios as follows:

$$p(p_1|\lambda_{k-1}) = p(p_1|\lambda_{k-2}) \frac{\ell(X_{k-1})p_1 + (1 - p_1)}{\int \{\ell(X_{k-1})p_1 + (1 - p_1)\} p(p_1|\lambda_{k-2}) dp_1} \quad (2.27)$$

Thus (2.22) may be rewritten in the form

$$L(X_k|\lambda_{k-1}) = \frac{\int \ell(s_k) \frac{p_1}{1 - p_1} \{\ell(X_{k-1})p_1 + (1 - p_1)\} p(p_1|\lambda_{k-2}) dp_1}{\int \{\ell(X_{k-1})p_1 + (1 - p_1)\} p(p_1|\lambda_{k-2}) dp_1} \quad (2.28)$$

and a system may be synthesized in the form of Fig. 7.

The solution when other parameters are also unknown is very similar since in this case we have the basic equation:

$$L(X_k|\lambda_{k-1}) = \iint \iota(X_k|\theta) \frac{p_1}{1 - p_1} p(p_1, \theta|\lambda_{k-1}) dp_1 d\theta \quad (2.29)$$

But $p(\theta, p_1 | \lambda_{k-1})$ may be written as

$$p(\theta, p_1 | \lambda_{k-1}) = p(\theta | p_1, \lambda_{k-1}) p(p_1 | \lambda_{k-1}) \quad (2.30)$$

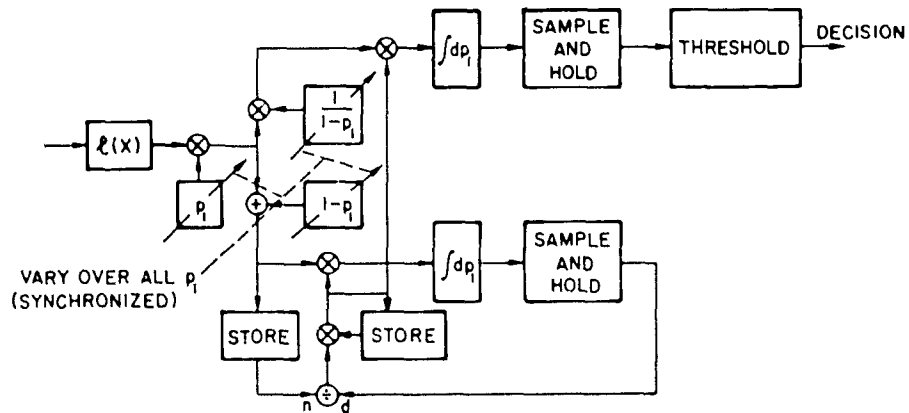
and its computation performed by the systems of Fig. 3. Therefore by writing

$$L(X_k | p_1, \lambda_{k-1}) = \int \ell(X_k | \theta) p(\theta | p_1, \lambda_{k-1}) d\theta \quad (2.31)$$

(which is computed by the system of Fig. 3 with a suitable choice of α), Eq. (2.29) becomes

$$L(X_k | \lambda_{k-1}) = \int L(X_k | p_1, \lambda_{k-1}) \frac{p_1}{1 - p_1} p(p_1 | \lambda_{k-1}) dp_1 \quad (2.32)$$

which is functionally very similar to (2.22). Thus the system of Fig. 7 with $\ell(X)$ replaced by $L(X_k | p_1, \lambda_{k-1})$ is the required system.



33362

FIG. 7. A SYSTEM DESIGNED TO LEARN THE A PRIORI PROBABILITIES.

F. EXTENSION TO THE MULTIPLE-HYPOTHESIS DECISION PROBLEM

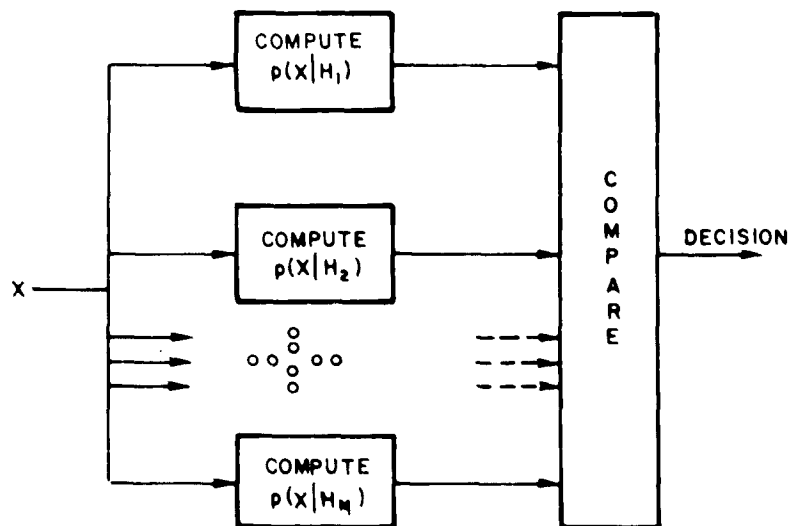
In the multiple-hypothesis problem we are given M possible classes into which we must categorize the vector X . There are $(M-1)$ possible errors associated with each of the M classes. The Bayes optimum solution depends upon the M^2 different weights which may be assigned to each error; that is, a general solution requires the comparison of weighted a posteriori probabilities

$$p(X|H_i) \quad i = 1, 2, \dots, M$$

where

H_i = hypothesis that X is in class i

A general form of the optimum system is shown in Fig. 8. From this solution it can be seen that in multiple-hypothesis testing the conditional probability $p(X|H_i)$ plays the same role as the likelihood ratio plays in the binary detection problem. In order to obtain a learning solution



33378

FIG. 8. A MULTIPLE-HYPOTHESIS MACHINE.

we assume that each class is characterized by an unknown vector A_i , and apply the same reasoning that was applied to the signal-detection problem. (We assume that the A_i are independent, that the X_i are conditionally independent, and that the $p(H_i)$ are known.) That is,

$$p(X_1, \dots, X_k | A_i, H_i) = p(X_1 | A_i, H_i) p(X_2 | A_i, H_i) \dots p(X_k | A_i, H_i)$$

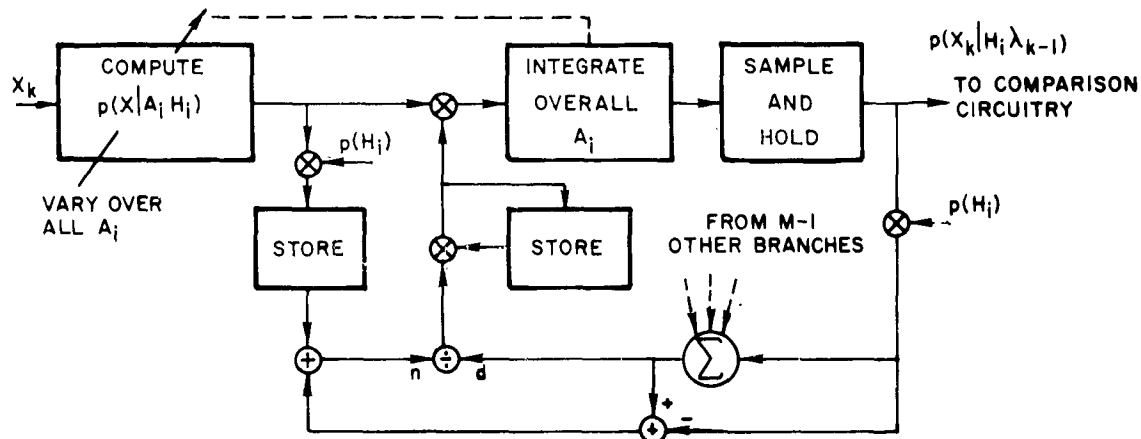
$$p(A_1, \dots, A_M) = p(A_1) p(A_2) \dots p(A_M)$$

The solution is given by

$$p(X_k | H_i, \lambda_{k-1}) = \int p(X_k | H_i, A_i) \cdot \frac{\left\{ p(X_{k-1} | H_i, A_i) p(H_i) + \sum_{j \neq i}^M p(X_{k-1} | H_j, \lambda_{k-2}) p(H_j) \right\}}{\sum_{j=1}^M p(X_{k-1} | H_j, \lambda_{k-2}) p(H_j)} \cdot p(A_i | \lambda_{k-2}) dA_i \quad (2.33)$$

This equation shows the same recursive form shown by Eq. (2.12), and leads to a similar system as depicted in Fig. 9. Since this system computes only one of the required M conditional probabilities, there must be $(M-1)$ more systems that are identical except for initial probability distribution $p_o(A_i)$ and probability of occurrence $p(H_i)$. These will in general be different; but in the case where the $p(H_i)$ are all the same, the $p_o(A_i)$ must be different, or all computer branches will "learn" the same thing, and the system as a whole will learn nothing.

It is interesting to note that Eq. (2.33) verifies our intuitive feeling that if we do not have some initial knowledge that the patterns to be recognized are somehow different, we are not able to learn without some external aid.



33379

FIG. 9. A MULTIPLE-HYPOTHESIS MACHINE WHICH LEARNS WITHOUT A TEACHER.

G. SUMMARY OF CHAPTER II

In this chapter we have mathematically described a class of decision problems in which the pertinent probability measures are known except for some set of fixed parameters. We have developed a class of systems which will solve such problems when a sequence of "learning" observations is available that contains information about the unknown parameter. This class of systems may take the form of either the "sequential" or the "parallel" canonical systems of Figs. 2 or 3. The resulting systems are optimum at each decision instant in the sense that of all possible systems based on the same a priori information and utilizing the same set of observations, these systems will provide the minimum average risk decision. The systems are also fixed in size for arbitrary learning sequences. The optimality and fixed size represent important advantages over both conventional and prior learning systems.

III. PERFORMANCE OF LEARNING SYSTEMS

It is the purpose of this chapter to investigate techniques for determining bounds on the performance of the previously developed learning system in specific cases, and to gain some insight into the way in which this performance depends upon the number of samples in the "learning" set.

A. PERFORMANCE MEASURES

This chapter is concerned with two aspects of system performance. The first is the average risk associated with a decision, the second is the rate at which the system converges to the optimum system given a priori knowledge of the parameter.

The average risk for the binary decision problem [Ref. 3] may be defined as

$$\rho = p_1 P_I + L p_2 P_{II} \quad (3.1)$$

where p_1 = a priori probability of hypothesis 1 being true
 $p_2 = 1 - p_1$ = a priori probability of hypothesis 2 being true
 P_I = probability of deciding that H_2 is true when H_1 is actually true
 P_{II} = probability of deciding that H_1 is true when H_2 is actually true
 L = cost of a type II error relative to a type I error

It will be convenient to discuss performance in terms of the signal-detection problem. In this case P_I becomes the probability of a miss P_M and P_{II} becomes the probability of false alarm P_{FA} (H_1 = signal present, H_2 = signal absent).

The rate of system convergence may be measured in terms of the decrease of the difference between transient average risk and steady-state average risk as a function of the number of observations (k). This measure will be called the "risk error" and defined as

$$\epsilon_{\rho} = \rho(d^*(\lambda_k)) - \rho(d^*(\hat{\theta})) \quad (3.2)$$

where $\rho(d^*(\lambda_k))$ = average risk of the system in the transient state
after k learning observations
 $\rho(d^*(\hat{\theta}))$ = average risk of the system given a priori knowledge
of θ

B. A TECHNIQUE FOR BOUNDING THE PERFORMANCE

Although there are several possible approaches to the evaluation of performance, only one will be presented. This approach is applicable to the class of learning problems restricted to those which may be expressed in terms of the detection of one of a finite set of signals embedded in noise. If the noise is additive, white, and normally distributed and if the signals are orthogonal, this approach results in some remarkably simple results which are in good agreement with intuition. For more general problems, the results are so dependent upon the particular problem that no useful or enlightening information has been uncovered.

Assume we have synthesized a system to detect the presence of a signal of unknown waveform which must be drawn from a set of m signals $\{S_1, S_2, \dots, S_m\}$ depending on a discrete set of parameters $\{\theta_1, \theta_2, \dots, \theta_m\}$; i.e., $S_i = S(\theta_i)$. Let the signals be embedded in noise. Then the optimum learning system will compute

$$L(X|\lambda_k) = \sum_{i=1}^m P(e_i|\lambda_k) L(X|\theta_i) \quad (3.3)$$

and compare it to a threshold $\gamma = Lp_2/p_1$. Let

- $\hat{\theta}$ = true value of θ
- $d^*(\lambda_k)$ = the Bayes decision rule based on λ_k
- $d^*(\hat{\theta})$ = the Bayes decision rule given knowledge of $\hat{\theta}$
- $d'(\lambda_k)$ = any non-Bayes decision rule based on λ_k
- $r(d)$ = average risk associated with decision rule d

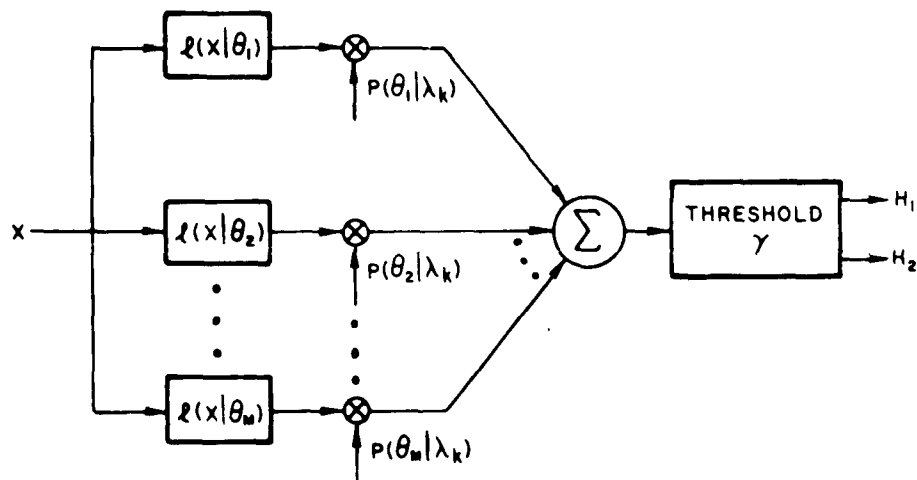
The average risk after k observations will be greater than the risk given that the $\hat{\theta}$ were known:

$$\rho(d^*(\lambda_k)) \geq \rho(d^*(\hat{\theta})) \quad (3.4)$$

The risk will be less than the risk of any other system based on λ_k :

$$\rho(d^*(\lambda_k)) \leq \rho(d'(\lambda_k)) \quad (3.5)$$

These properties follow from the Bayes nature of the decision rule. The optimum system is sketched for convenience in Fig. 10.



33380

FIG. 10. AN OPTIMAL LEARNING SYSTEM.

To evaluate a bound, consider a suboptimum system which computes $P(\theta_1|\lambda_k)$ and $\bar{x}(x|\theta_1)$. Let this system determine the θ_1 for which $P(\theta_1|\lambda_k)$ is largest and compare the corresponding $\bar{x}(x|\theta_1)$ with the threshold γ to make a decision.[†] If $P(\hat{\theta}|\lambda_k)$ is greater than $1/2$

[†] This suboptimum system is closely related to the "maximum-likelihood receiver" of Helstrom [Ref. 22, p. 238] and to the "type III" receiver of Wainstein and Zubakov [Ref. 23, p. 297]; however it is different in that it takes the past into account.

it will be largest, and the suboptimum system will have the same average risk as the system based on knowledge of $\hat{\theta}$. If $P(\hat{\theta}|\lambda_k) \leq 1/2$, it may not be the largest, and an incorrect likelihood computer may be chosen.

If the wrong θ_i is chosen, the risk may be bounded as follows. Define $P_{FA}^{(i)}(\hat{\theta})$ and $P_M^{(i)}(\hat{\theta})$ as

$$P_{FA}^{(i)}(\hat{\theta}) = \Pr \{ \ell(X|\theta_i) > \gamma | H_2, \hat{\theta} \} \quad \theta_i \neq \hat{\theta} \quad (3.6a)$$

$$P_M^{(i)}(\hat{\theta}) = \Pr \{ \ell(X|\theta_i) < \gamma | H_1, \hat{\theta} \} \quad \theta_i \neq \hat{\theta} \quad (3.6b)$$

where $\hat{P}_{FA} = \Pr \{ \ell(X|\hat{\theta}) > \gamma | H_2, \hat{\theta} \}$ is the probability of false alarm when $\hat{\theta}$ is known. Let

$$\epsilon_{FA}(\hat{\theta}) = \max_{i=1, \dots, Q} P_{FA}^{(i)}(\hat{\theta}) - \hat{P}_{FA} \quad (3.6c)$$

$$\epsilon_M(\hat{\theta}) = \max_{i=1, \dots, Q} P_M^{(i)}(\hat{\theta}) - 1 + \hat{P}_{FA} \quad (3.6d)$$

These two latter quantities are small in many interesting problems. For example, $\epsilon_{FA}(\hat{\theta}) = 0$ whenever the distribution of $\ell(X|\theta_i)$ conditioned on H_2 and $\hat{\theta}$ is independent of θ_i , as is the case in the detection of a set of signals in additive normal noise when the signals are orthogonal after whitening (i.e., let K be the covariance matrix of the noise), then the whitened signals are orthogonal if $S(\theta_i)_t K^{-1} S(\theta_j) = \delta_{ij} R$, where $\delta_{ij} = 1$ if $i = j$, and $\delta_{ij} = 0$ if $i \neq j$. In this particular case we may write

$$\ell(X|\theta_i) = \exp \left[-\frac{1}{2} R + S(\theta_i)_t K^{-1} X \right]$$

But since we may replace X by

$$S(\hat{\theta}) + N \quad \text{when } H_1 \text{ is true}$$

and by

N when H_2 is true

then the orthogonality of $S(\theta_1)$ and $S(\hat{\theta})$ allows us to write, for both H_1 and H_2 ,

$$\ell(X|\theta_1) = \exp \left[-\frac{1}{2} R + S(\theta_1)_t K^{-1} N \right] \quad (3.7)$$

Thus the distribution of $\ell(X|\theta_1)$ will be the same whether H_1 or H_2 is true, and $\epsilon_M(\hat{\theta})$ will be zero.

Next, we define two events, A_k and B_k as:

$$A_k = \text{event } P(\hat{\theta}|\lambda_k) > P(\theta_1|\lambda_k) \quad \text{for all } \theta_1 \neq \hat{\theta}$$

$$B_k = \text{event } P(\hat{\theta}|\lambda_k) \leq P(\theta_1|\lambda_k) \quad \text{for some } \theta_1 \neq \hat{\theta}$$

The risk of the suboptimum system, when $\hat{\theta}$ is true, is bounded by

$$\rho(d'(\lambda_k)|\hat{\theta}) \leq \rho(d^*(\hat{\theta})|\hat{\theta}) P(A_k) + P(B_k) \max_i \left[p_1 P_M^{(i)}(\hat{\theta}) + L p_2 P_{FA}^{(i)}(\hat{\theta}) \right] \quad (3.8)$$

Inserting the definition of $\epsilon_M(\hat{\theta})$, $\epsilon_{FA}(\hat{\theta})$, and $\rho(d^*(\hat{\theta})|\hat{\theta})$ in this expression gives

$$\begin{aligned} \rho(d'(\lambda_k)|\hat{\theta}) &\leq p_1 \hat{P}_M P(A_k) + L p_2 \hat{P}_{FA} [P(A_k) + P(B_k)] \\ &\quad + P(B_k) \{ p_1 (1 + \epsilon_M) + L p_2 \epsilon_{FA} - \hat{P}_{FA} \} \end{aligned} \quad (3.9)$$

where

$$\hat{P}_M = \Pr \{ \ell(X|\hat{\theta}) < \gamma | H_1, \hat{\theta} \}$$

is the miss probability when $\hat{\theta}$ is known, and the dependence of ϵ_M and ϵ_{FA} on $\hat{\theta}$ has been suppressed.

For most problems we are interested in the region where $\hat{P}_M$ and $\hat{P}_{FA}$ are small compared to 1, and $P(A_k)$ is nearly one, so that by combining (3.5) and (3.9) we obtain

$$\rho(d^*(\lambda_k)|\hat{\theta}) \leq \rho(d^*(\hat{\theta})|\hat{\theta}) + p_1 P(B_k) + p_1 \epsilon_M + L p_2 \epsilon_{FA} \quad (3.10)$$

Thus we may identify the risk error ϵ_ρ as

$$\epsilon_\rho = p_1 P(B_k) + p_1 \epsilon_M + L p_2 \epsilon_{FA} \quad (3.11).$$

Because the system performance is dependent on $\hat{\theta}$ through $P(B_k)$, $\epsilon_M(\hat{\theta})$, etc., the application of this bound to any particular problem is difficult; and as the performance becomes more and more dependent on $\hat{\theta}$, the bound becomes less useful because it is more and more difficult to obtain an evaluation of $P(B_k)$. To obtain some insight into the nature of this bound, let us evaluate the bound for the problem of detection of a signal embedded in additive white gaussian noise when the signal waveform is unknown. The signal may take on one of m orthogonal waveforms. As noted previously, the normality of the noise and the orthogonality of the signals insure that when the signal is not present the choice of the proper θ_i does not affect false-alarm probability; that is,

$$P_{FA}^{(i)}(\hat{\theta}) = \hat{P}_{FA} \quad \text{for all } \theta_i$$

so that $\epsilon_{FA}(\hat{\theta}) = 0$. Similarly, the normality of the noise and orthogonality of the signals insure that

$$P_M^{(i)}(\hat{\theta}) = 1 - P_{FA}^{(i)}(\hat{\theta}) \quad \text{for all } \theta_i + \hat{\theta}$$

so that $\epsilon_M(\hat{\theta}) = 0$. Thus we have

$$\epsilon_\rho = p_1 P(B_k) \quad (3.12)$$

In Appendix A we apply a Tchebysheff-type bound to show that $p_1 P(B_k)$ (hence ϵ_ρ) is bounded as

$$\epsilon_\rho = p_1 P(B_k) \leq \frac{4(1 + p_1 p_2 R)}{p_1 k R - 4 \left[\frac{m-1}{m} k R (1 + p_1 p_2 R) \right]^{1/2}} \quad (3.13)$$

where

$$R = \frac{E\{S(\theta_i)_t S(\theta_i)\}}{\sigma_n^2} \quad \text{is a constant signal-to-noise ratio}$$

m = number of orthogonal signals

p_1 = probability of signal occurrence

$p_2 = 1 - p_1$ = probability no signal will occur

k = number of observations in the learning sequence

For large k this bound is asymptotic to

$$\frac{4(1 + p_1 p_2 R)}{p_1 k R}$$

Thus the system performance converges to the performance of the system which has a priori knowledge of the signal waveform at least as fast as inversely with $p_1 k$, the average number of learning samples which contain a signal in a sequence of length k .

C. EXAMPLE

The problem of detecting a signal of unknown frequency in white gaussian noise, which was used as example 2 of Chapter II, is an example

of a problem in which the system performance may be evaluated by the preceding procedure. In this example, the set of possible signals is

$$s_i = \begin{cases} a \cos (\omega_i t + \phi) & 0 \leq t \leq T \\ 0 & \text{elsewhere} \end{cases} \quad (3.14)$$

where ϕ = a uniformly distributed random variable

a = a Rayleigh-distributed random variable with parameter A^2

In this case all of the preceding conditions are met. We identify

$$R = \frac{A^2}{2N_o W} \quad (3.15)$$

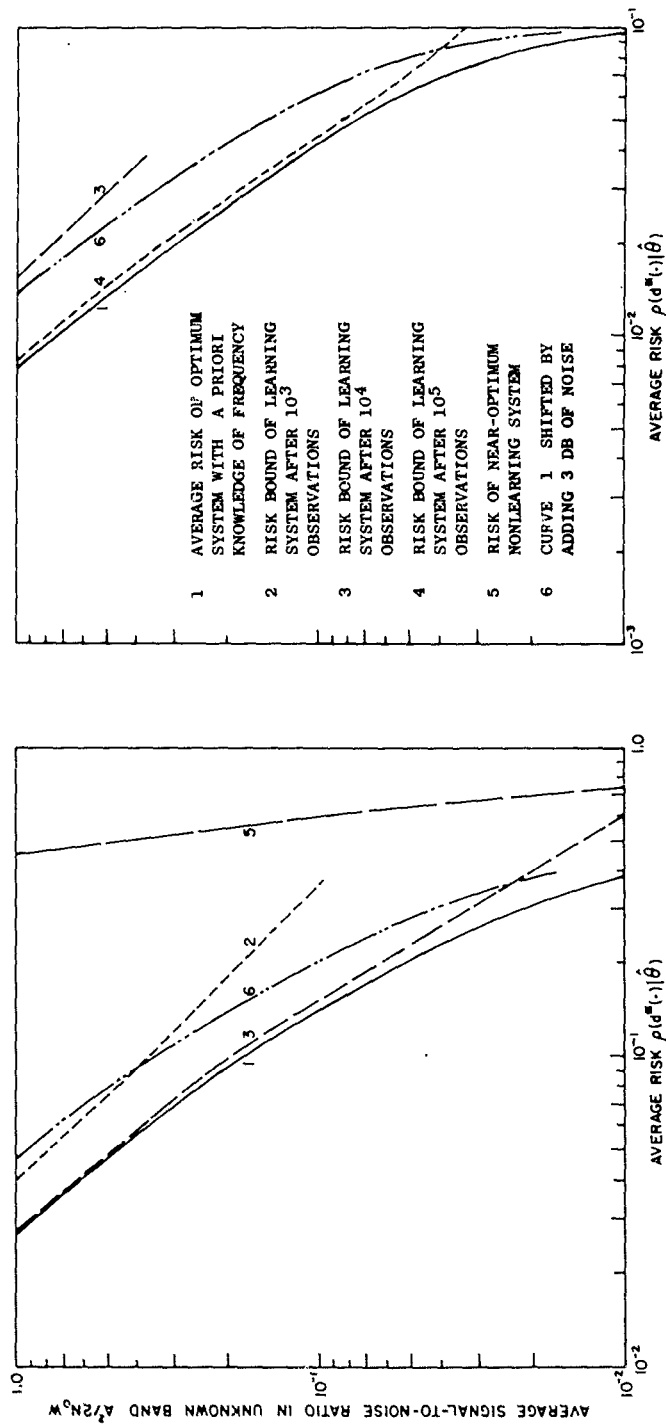
where $N_o/2$ = noise spectral density

W = total band to be searched

and we can compute $\rho(d^*(\hat{S})|\hat{S})$ using standard procedures [Ref. 23, p. 173ff], and evaluate ϵ_ρ using Eq. (3.13). Parts (a) and (b) of Fig. 11 show the results for the case where there are 200 possible frequencies within the band W . Also shown are the following:

1. The performance of the optimum system given a priori knowledge of $\hat{S}$ for $L = 1$, and for $p_1 = 1/2$ as a function of the signal-to-noise ratio, R .
2. The performance curve of $\rho(d^*(\hat{S})|\hat{S})$ shifted by 3 db on the R axis.
3. Bounds for the performance of the learning receiver for 1,000 and 10,000 samples.
4. The performance of a near-optimum, 200-channel receiver (Wainstein and Zubakov "type III" receiver [Ref. 23, p. 300ff]) which does not learn.

It is clear that for case 3 above, the incremental risk introduced by lack of knowledge of the signal frequency is very small after 10,000 learning observations, and that it is not much different--after 1,000



b. $L = 1, p_1 = 1/10$

a. $L = 1, p_1 = 1/2$

FIG. 11. PERFORMANCE BOUNDS FOR A LEARNING SYSTEM.

33361

observations--from the incremental risk introduced by doubling the noise power when the frequency is known. It is also clear that for almost any task the nonlearning receiver would be virtually useless at the signal-to-noise ratios shown.

D. OTHER TECHNIQUES TO OBTAIN PERFORMANCE BOUNDS

A second technique to obtain bounds on system performance may be based on the use of Chernoff bounds for the tail probability of a sum of random variables. This technique is applicable to Bayes optimum systems which learn either with or without a teacher, since they may both be described by

$$\ell_k = \ell(X_k | \lambda_{k-1}) = \int \ell(X_k | \theta) p(\theta | \lambda_{k-1}) d\theta \quad (3.16)$$

This sequence of likelihoods $\{\ell_i; i = 1, 2, \dots\}$ is a martingale sequence as shown in Appendix C. It may be centered at its expectation by considering the "gain" at each new observation. Let $y_i = \ell_i - \ell_{i-1}$, then

$$\ell_k = \sum_{i=1}^k y_i$$

Shannon [Ref. 27] has applied Chernoff's bounding technique to such martingale sequences, and his work is almost directly applicable in this case.

Bounds on the tail probabilities may be written as inequalities involving bounds on the semi-invariant generating functions for the martingale sequence. For example, we may let $\nu_k(u | H_j)$ be the moment-generating function for ℓ_k conditioned on H_j being true for X_k (H_1 = hypothesis that signal is present; H_2 = hypothesis that signal is absent). That is,

$$\nu_k(u | H_j) = \int \dots \int \exp(u \sum y_i) dP(y_i, y_{i-1}, \dots, y_1 | H_j) \quad (3.17)$$

Now define bounding functions by the relations

$$\gamma_i(u) \cong \int \exp(uy_i) dP(y_i | y_{i-1}, \dots, y_1) \quad (3.18a)$$

$$\gamma_k(u | H_j) \cong \int \exp(uy_k) dP(y_k | y_{k-1}, \dots, y_1, H_j) \quad (3.18b)$$

and

$$\mu_i(u) = \log \gamma_i(u) \quad (3.19a)$$

$$\mu_k(u | H_j) = \log \gamma_k(u | H_j) \quad (3.19b)$$

Suppose that we can find a single bound for all $i \leq k-1$, and call this $\mu_o(u)$. Then we can show that, for some $a, b > 0$, real:

$$\begin{aligned} \Pr \{ \ell_k \geq (k-1)\mu_o'(u) + \mu_k'(u | H_j) \} \leq \exp \{ (k-1)[\mu_o(u) - u\mu_o'(u)] \\ + \mu_k(u | H_j) - u\mu_k'(u | H_j) \} \end{aligned} \quad (3.20a)$$

for $0 \leq u \leq b$, and

$$\begin{aligned} \Pr \{ \ell_k \leq (k-1)\mu_o'(u) + \mu_k(u | H_j) \} \leq \exp \{ (k-1)[\mu_o(u) - u\mu_o'(u)] \\ + \mu_k(u | H_j) - u\mu_k'(u | H_j) \} \end{aligned} \quad (3.20b)$$

for $-a \leq u \leq 0$.

One technique for finding a suitable $\mu_o(u)$ is to find two cumulative distribution functions $\phi_1(y)$ and $\phi_2(y)$ which bound $P(y_i | y_{i-1}, \dots, y_1)$ above and below for all y_i . We then choose $\phi_o(y)$ such that

$$\begin{aligned}
(1) \quad \phi_o(y) &= \phi_1(y) & y \leq \alpha \\
(2) \quad \phi_o(y) &= \phi_2(y) & y \geq \beta \\
(3) \quad \phi_o(y) &= \phi_1(\alpha) = \phi_2(\beta) & \alpha \leq y \leq \beta \\
(4) \quad \int y \, d\phi_o(y) &= 0
\end{aligned} \tag{3.21}$$

Define

$$\mu_o(u) = \log \int e^{uy} \, d\phi_o(y) \tag{3.22}$$

Then $\mu_o(u)$ is a bound of the desired type. We can use this same approach to bound $\mu_k(u|H_j)$ by conditioning $\phi_1(y|H_j)$ and $\phi_2(y|H_j)$ on H_j . Unfortunately at this point the technique requires specification of the particular problem in more detail; i.e., the distributions of $\lim_{k \rightarrow \infty} \hat{\ell}_k$ and ℓ_o must be specified. Although this is in general possible, for the case of detection of an unknown signal in gaussian noise, both distributions are log-normal and the moment-generating functions do not exist for any u interval. This problem may be overcome by noting that any practical system to compute $\hat{\ell}_k$ has a finite dynamic range, and by truncating the distribution at this limit. Such truncation makes evaluation of the bounds very difficult. Numerical solutions may of course be found for any particular problem by means of a computer solution; however the results can only be expressed numerically and will most likely shed little light on the question of performance in general.

There are two other methods for determining system performance which should be considered by anyone setting out to decide whether or not a learning system is worth the cost in time, complexity, and money for any particular problem. These methods involve either a direct evaluation of the cumulative probability distribution of $i(X|\lambda_k)$ from the known statistics of X and λ_k , or a determination of the statistics of $i(X|\lambda_k)$ by simulation of the system and design of an experiment to determine the desired performance measures. Both of these approaches seem to

be complex for reasonably large k ; however in certain problems, particularly where convergence is rapid, either approach may be useful.

IV. LEARNING TIME-VARYING PARAMETERS

In the previous chapters a technique has been developed which will allow the synthesis of systems which learn without a teacher when the unknown parameter is fixed. This technique was accomplished by treating the unknown parameter as a random variable. In this chapter the same problem will be examined for the case where the unknown parameter is not fixed, but varies with time. A synthesis technique for systems to solve this problem will be developed by taking an approach similar to the previous one and treating the unknown parameter as a random variable which is time varying.

A. MODELS FOR THE TIME-VARYING PARAMETER PROBLEM

As in Chapter II the problem considered will be the binary decision problem phrased in terms of detection of a signal which depends upon a set of unknown parameters. The results may be generalized to obtain the extension to multiple-hypothesis testing as in Chapter II.

The data to be used consist of an observation X_k and a learning sequence $\{\lambda_{k-1} = X_1, X_2, \dots, X_{k-1}\}$. Each observation contains a signal corrupted by noise, or it contains noise alone, and it is desired to synthesize a system to decide whether or not the k^{th} observation (X_k) contains a signal. The Bayes-optimum system making optimal use of the learning sequence is required. This problem differs from the problem of Chapter II in this sense: the values of the unknown signal parameters are not the same from observation to observation. This fact is indicated by indexing the parameter set with a lowercase letter "c"; i.e., the signal parameters defining the signal present (if any) in the current observation (X_k) are designated θ_c .

Formally, we let

$$H_1 = \text{hypothesis that } X_k = S(\theta_c) \oplus N$$

$$H_2 = \text{hypothesis that } X_k = N$$

where $S(\theta_c)$ is the current signal vector (unknown parameters) and N is the noise vector.

For a particular problem the statistical nature of the noise and the corrupting operation are assumed to be known so that the only unknowns are the signal parameters. In order to solve the problem a statistical model of the signal-parameter variations from observation to observation is required. The statistical model must include a description of the way in which the current values of the parameters depend upon past values, and a description of the statistics of the times of occurrence of changes.[†] The former description will be called "value dependence" and the latter "time dependence."

The value dependence of the signal parameters may be described by the probability density of the c^{th} realization of the signal conditioned on all of the past realizations:

$$p(\theta_c | \theta_{c-1}, \theta_{c-2}, \dots, \theta_1)$$

In some problems, particularly the frequency-hopping signal reconnaissance problem to be described in example 2, the c^{th} realization will be independent of the past, so that

$$p(\theta_c | \theta_{c-1}, \theta_{c-2}, \dots, \theta_1) = p_o(\theta_c) \quad (4.1)$$

In other problems the dependence may be Markov so that

$$p(\theta_c | \theta_{c-1}, \theta_{c-2}, \dots, \theta_1) = p(\theta_c | \theta_{c-1}) \quad (4.2)$$

In yet other problems the entire past may enter; however, these problems lead to systems which grow in size with k . For this reason the value dependence will be restricted to be at worst M^{th} -order Markov.

[†]Throughout this chapter it is assumed that changes in parameter value can take place only at a (countable) set of discrete instants in time.

The time dependence cannot be described as generally as the value dependence; however there are two types of time dependence which are of particular interest because they occur frequently in physical problems. The first type will be designated the "general random walk," since we assume that a change takes place at the start of each observation. The amount of change, as well as the direction, depends on the past history and is described by $p(\theta_c | \theta_{c-1}, \theta_{c-2}, \dots, \theta_{c-M})$. An example of an unknown time-varying parameter which may be approximated by this model is the complex gain of a communication channel which is slowly varying with respect to the duration of one signal (see example 1 of this chapter).

The second type of time dependence will be designated a "binomial" dependence. In this model the changes in the parameter occur at moments which coincide with the start of an observation, but changes do not occur at each new observation. The probability that n changes will occur in j trials is the binomial distribution,

$$P_n(j) = \binom{n}{j} p^j (1 - p)^{n-j} \quad (4.3)$$

where p is the probability of a change in one trial. Once again the value dependence is described by the conditional density $p(\theta_c | \theta_{c-1}, \theta_{c-2}, \dots, \theta_{c-M})$. An example of a parameter which has a "binomial" time dependence is the frequency of a frequency-hopping signal as explained in example 2 of this chapter.

B. SOLUTION TO THE PROBLEM

In order to obtain a solution to the learning problem, θ_c is treated as a random variable and the a posteriori distribution of θ_c is learned. As before, the Bayes system will compute the conditional likelihood ratio

$$t(x_k | \lambda_{k-1}) = \int t(x_k | \theta_c) p(\theta_c | \lambda_{k-1}) d\theta_c \quad (4.4)$$

and compare it to the appropriate threshold. Since it is assumed that $\ell(X_k | \theta_c)$ has known form, the problem again reduces to the computation of $p(\theta_c | \lambda_{k-1})$. In order to compute this function, a time-dependence model must be given; hence the problem may be treated for two cases:

1. Case 1, General-Random-Walk Time Dependence

a. General Solution

In this case the index c will coincide with k since θ will change with each new observation; therefore,

$$p(\theta_k | \lambda_{k-1}) = \frac{\int \dots \int \left\{ \prod_{i=1}^k p(\theta_i | \theta_{i-1}, \dots, \theta_1) \right\} \left\{ \prod_{i=1}^{k-1} p(X_i | \theta_i) \right\} d\theta_{k-1} \dots d\theta_1}{p(\lambda_{k-1})} \quad (4.5)$$

Thus if the value dependence is not at least as simple as M^{th} -order Markov [i.e., if $p(\theta_k | \theta_{k-1}, \dots, \theta_1)$ may not be written as $p(\theta_k | \theta_{k-1}, \dots, \theta_{k-M})$], then the system to compute $p(\theta_k | \lambda_{k-1})$ must grow in size linearly with k . The complexity of the system would grow much more rapidly. This is the reason for restriction of the statistics describing the value dependence of the parameter to be M^{th} -order Markov with M finite.

b. First-Order Markov Value Dependence--Vector Parameters

For simplicity in obtaining a system from this equation, assume that the value dependence of the parameter is first-order Markov so that

$$\begin{aligned} p(\theta_k | \lambda_{k-1}) &= \int p(\theta_k | \theta_{k-1}) p(\theta_{k-1} | \lambda_{k-1}) d\theta_{k-1} \\ &= \int p(\theta_k | \theta_{k-1}) \frac{p(X_{k-1} | \theta_{k-1})}{p(X_{k-1} | \lambda_{k-2})} p(\theta_{k-1} | \lambda_{k-2}) d\theta_{k-1} \end{aligned} \quad (4.6)$$

Equation (4.6) illustrates once again the recursive nature of the computations; that is, once $p(\theta_{k-1}|\lambda_{k-2})$, $p(x_{k-1}|\theta_{k-1})$, and $p(x_{k-1}|\lambda_{k-2})$ are computed (all of these quantities will be computed during the previous observation-decision cycle), $p(\theta_k|\lambda_{k-1})$ can be computed.

Equation (4.6) may be rewritten in terms of likelihood ratios as follows:

$$p(\theta_k|\lambda_{k-1}) = \int p(\theta_k|\theta_{k-1}) \left[\frac{\ell(x_{k-1}|\theta_{k-1}) + \alpha}{\ell(x_{k-1}|\lambda_{k-2}) + \alpha} \right] p(\theta_{k-1}|\lambda_{k-2}) d\theta_{k-1} \quad (4.7)$$

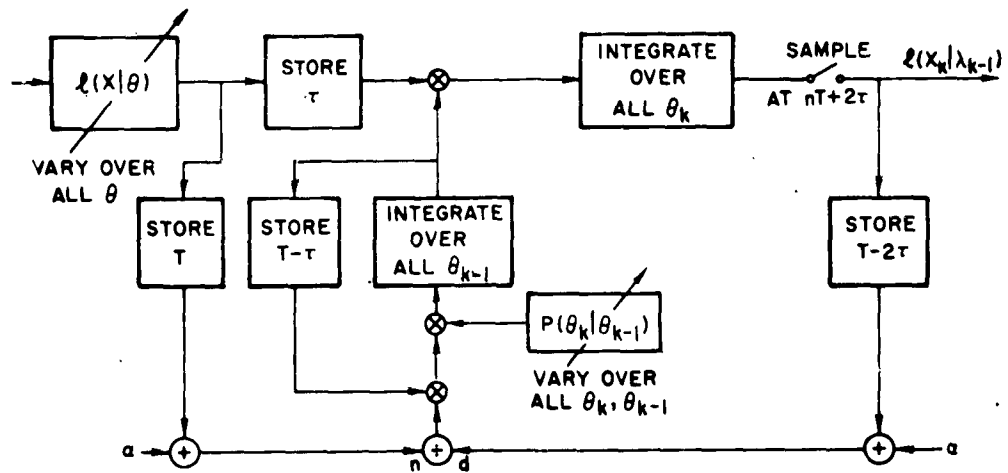
where as before $\alpha = p(H_1)/p(H_2)$. After rewriting Eq. (4.4) as shown below, the required system may be conveniently synthesized.

$$\begin{aligned} \ell(x_k|\lambda_{k-1}) = & \int \ell(x_k|\theta_k) \int p(\theta_k|\theta_{k-1}) \left[\frac{\ell(x_{k-1}|\theta_{k-1}) + \alpha}{\ell(x_{k-1}|\lambda_{k-2}) + \alpha} \right] \\ & \cdot p(\theta_{k-1}|\lambda_{k-2}) d\theta_{k-1} d\theta_k \end{aligned} \quad (4.8)$$

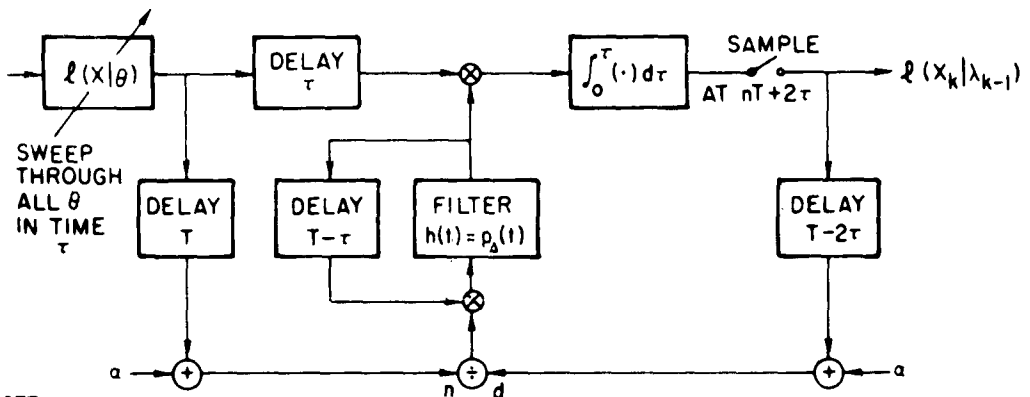
From (4.7) and (4.8) the system shown in Fig. 12a may be synthesized. This system operates in a manner similar to that described in Chapter II. It performs three operations:

1. Compute $\ell(x_k|\theta_k)$ for each possible θ
2. Compute $p(\theta_k|\lambda_{k-1})$ for each possible θ
3. Weight (1) by (2), and sum over all θ .

It is in the performance of the second operation that the time variation of θ is taken into account by including the three components [the $p(\theta_k|\theta_{k-1})$ generator, the multiplier, and the integrator] in the probability loop. The $\ell(x|\theta)$ computer must "sweep" through all values of θ , and the $p(\theta_k|\theta_{k-1})$ generator must "sweep" through all combinations of values of θ_k and θ_{k-1} . If there are only a finite number of values of θ , the system may be realized in a parallel form by utilizing a set of parallel computers $\{\ell(x|\theta^{(i)})\}; i = 1, 2, \dots\}$, where i indexes the possible values of θ . In this case the two integrators are replaced by



a. Case 1b



33377

T = TIME OF ONE OBSERVATION
 tau = TIME TO SWEEP THROUGH ALL theta
 ASSUME ALL theta_k, theta_{k-1} SWEEPED IN 2tau < T

b. Case 1c

FIG. 12. LEARNING SYSTEM FOR GENERAL-RANDOM-WALK TIME DEPENDENCE.

summers with inputs from the parallel circuitry. A block diagram of such a general parallel system is difficult to draw; however the parallel form is used in the solution of example 2 later in this chapter.

c. First-Order Markov Value Dependence--Scalar Parameters

If the unknown parameter is scalar and has a random-walk type of dependence on the past, the system of Fig. 12a simplifies somewhat.

In this case, we may represent θ_k as a perturbation of θ_{k-1} ; i.e., let

$$\theta_k = \theta_{k-1} + \Delta_k \quad (4.9)$$

where Δ_k is independent of θ_{k-1} . Let the distribution of Δ_k be $p_\Delta(z)$. Then a simple transformation will provide Eq. (4.10).

$$p(\theta_k | \theta_{k-1}) = p_\Delta(\theta_k - \theta_{k-1}) \quad (4.10)$$

Equations (4.7) and (4.8) may be rewritten as

$$p(\theta_k | \lambda_{k-1}) = \int p_\Delta(\theta_k - \theta_{k-1}) \left[\frac{\ell(x_{k-1} | \theta_{k-1}) + \alpha}{\ell(x_{k-1} | \lambda_{k-2}) + \alpha} \right] \cdot p(\theta_{k-1} | \lambda_{k-2}) d\theta_{k-1} \quad (4.11)$$

$$\ell(x_k | \lambda_{k-1}) = \int \ell(x_k | \theta_k) \int p_\Delta(\theta_k - \theta_{k-1}) \left[\frac{\ell(x_{k-1} | \theta_{k-1}) + \alpha}{\ell(x_{k-1} | \lambda_{k-2}) + \alpha} \right] \cdot p(\theta_{k-1} | \lambda_{k-2}) d\theta_{k-1} d\theta_k \quad (4.12)$$

When θ is a scalar, the system can be realized by sweeping through the range of θ in some interval τ (which must be less than half the observation interval $T/2$ for real-time operation). In this case θ_k and θ_{k-1} are two different time variables, and Eq. (4.11) represents a convolution. For this reason, the system may be realized as shown in Fig. 12b, where the only difference from the system with fixed parameter is the filter with impulse response $p_\Delta(t)$ in the probability loop. In order to insure that this filter may be realized, the delay of τ in the forward loop has been added. This concept of the filter $p_\Delta(t)$ will be useful in the solution of the second example.

If the parameter value dependence is first-order Markov, but not representable as an independent perturbation, then $p(\theta_k | \theta_{k-1})$ will not lead to a time-invariant filter as above. Instead it will lead to a time-varying filter as a replacement for the filter with impulse response $p_{\Delta}(t)$ in Fig. 12b. The replacement will have a time-varying impulse response

$$h(t, \gamma) = p(\theta_k = t | \theta_{k-1} = \gamma) \quad (4.13)$$

The output of this filter at time t , for an input $z(t)$, is defined by

$$e_o(t) = \int_{-\infty}^{\infty} h(t, \gamma) z(\gamma) d\gamma \quad (4.14)$$

Methods for the realization of such filters are beyond the scope of this study; however one method for a particular form of $h(t, \gamma)$ is suggested in example 1. For other methods see Refs. 25, 26.

d. M^{th} -Order Markov Value Dependence--Vector Parameters

If the unknown parameter is M^{th} -order Markov, the same general approach to system synthesis may be taken. Equation (4.5) becomes

$$p(\theta_k | \lambda_{k-1}) = \int \dots \int \prod_{i=1}^M \left[\frac{p(X_{k-i} | \theta_{k-i})}{p(X_{k-i} | \lambda_{k-i})} p(\theta_{k-i} | \lambda_{k-i}) \right] \cdot p(\theta_k | \theta_{k-1}, \dots, \theta_{k-M}) d\theta_{k-1} \dots d\theta_{k-M} \quad (4.15)$$

The change which this requires in the block diagram is simple enough; however, the complexity of the system, even for scalar θ , rapidly becomes intolerable. To see that this is true, assume that θ is scalar and $M = 2$. Then somewhere within the system, the function $p(\theta_k | \theta_{k-1}, \theta_{k-2})$ must be stored for all possible combinations of θ_k ,

θ_{k-1} , and θ_{k-2} . If there are, say, N possible values which θ may take on, the storage required is N^3 . In general, the storage increases as N^M .

2. Case 2, Binomial Time Dependence

a. General

In the case of binomial time dependence the integral-valued variable j is defined as follows:

j = number of observations since the last change in θ

Then since the changes occur at moments which are binomially distributed, j will be exponentially distributed as follows:

$$P(j) = p(1-p)^{j-1} \quad (4.16)$$

where p is the probability of a change in θ .

In order to obtain a recursive relation from which a system may be synthesized, the distribution of θ is conditioned on j as well as on the last value of θ ; i.e.,

$$p(\theta_c | \lambda_{k-1}) = \sum_j P(j) p(\theta_c | j, \lambda_{k-1}) \quad (4.17)$$

But

$$p(\theta_c | j, \lambda_{k-1}) = \frac{p(\lambda_{k-1} | \theta_k, j)}{p(\lambda_{k-1} | j)} p(\theta_k | j) \quad (4.18a)$$

or

$$\begin{aligned}
p(\theta_c | j, \lambda_{k-1}) &= \frac{\prod_{i=1}^{k-1} p(X_i | \theta_c) p(\lambda_{k-j-1} | \theta_c, j) p(\theta_c | j)}{\prod_{i=1}^{k-j} p(X_i | \lambda_{i-1}) p(\lambda_{k-j-1} | j)} \\
&= \prod_{i=1}^{k-j} \left[\frac{p(X_i | \theta_c)}{p(X_i | \lambda_{i-1})} \right] p(\theta_c | \lambda_{k-j-1}, j) \\
&= \prod_{i=1}^{k-j} \frac{\ell(X_i | \theta_c) + \alpha}{\ell(X_i | \lambda_{i-1}) + \alpha} p(\theta_c | \lambda_{k-j-1}, j) \quad (4.18b)
\end{aligned}$$

Also,

$$p(\theta_c | \lambda_{k-j-1}, j) = \int p(\theta_c | \theta_{c-1}, \lambda_{k-j-1}, j) p(\theta_{c-1} | \lambda_{k-j-1}, j) d\theta_{c-1} \quad (4.19)$$

Hence (4.17) may be rewritten as

$$\begin{aligned}
p(\theta_c | \lambda_{k-1}) &= \sum_j P(j) \prod_{i=1}^{k-1} \left[\frac{\ell(X_i | \theta_c) + \alpha}{\ell(X_i | \lambda_{i-1}) + \alpha} \right] \int p(\theta_c | \theta_{c-1}, \lambda_{k-j-1}, j) \\
&\quad \cdot p(\theta_{c-1} | \lambda_{k-j-1}, j) d\theta_{c-1} \quad (4.20)
\end{aligned}$$

Note that $p(\theta_{c-1} | \lambda_{k-j-1}, j)$ is the value of $p(\theta_c | \lambda_{k-1})$ calculated j observations ago, so that Eq. (4.20) is in some sense recursive. This fact will be exploited in the following paragraphs.

b. First-Order Markov Value Dependence--Vector Parameters

In order to interpret Eq. (4.20) as a system block diagram, the problem is simplified by assuming that the value dependence of θ is first-order Markov. In this case

$$p(\theta_c | \lambda_{k-j-1}, j) = \int p(\theta_c | \theta_{c-1}) p(\theta_{c-1} | \lambda_{k-j-1}, j) d\theta_{c-1} \quad (4.21)$$

Call this P_{k-j-1} . In order to rewrite Eq. (4.20), denote

$$L_i = \frac{\ell(X_i | \theta_i) + \alpha}{\ell(X_i | \lambda_{i-1}) + \alpha} \quad (4.22)$$

Then Eq. (4.20) may be expanded and written as

$$p(\theta_k | \lambda_{k-1}) = P(1) L_{k-1} P_{k-2} \cdot \left\{ 1 + \frac{P(2)}{P(1)} \frac{P_{k-3}}{P_{k-2}} L_{k-2} \left[1 + \frac{P(3)}{P(2)} \frac{P_{k-4}}{P_{k-3}} L_{k-3} (1 + \dots) \right] \right\} \quad (4.23)$$

This function is recursive in the sense that once L_{k-1} and P_{k-2} are available, L_k and P_k may be computed from X_k and X_{k-1} . A system to realize this computation in delay-feedback form is shown in Fig. 13.

c. Independent Values--Vector Parameters

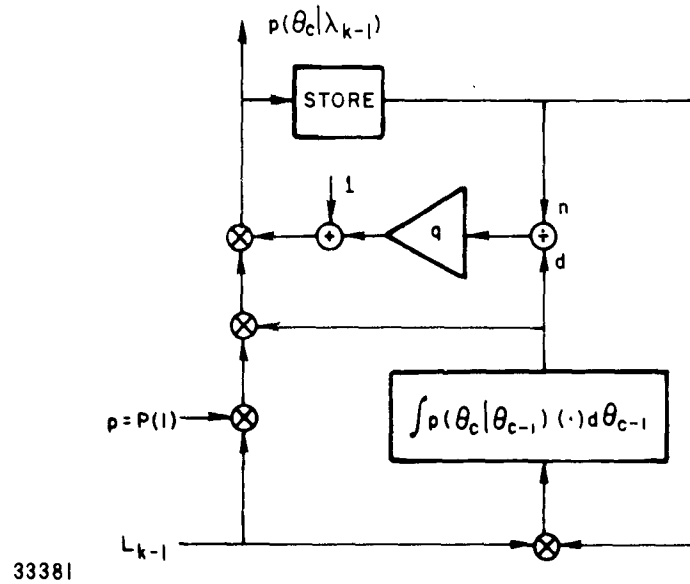
When the value of θ_c is independent of the past values of θ (which will occur, for example, in the frequency-hopping problem), Eq. (4.23) simplifies and the resulting system is more manageable. In this case,

$$p(\theta_c | \theta_{c-1}) = p_o(\theta_c) \quad (4.24)$$

hence

$$p(\theta_c | \lambda_{k-1}) = p_o(\theta_c) P(1) L_{k-1} \left\{ 1 + \frac{P(2)}{P(1)} L_{k-2} \left[1 + \frac{P(3)}{P(2)} L_{k-3} (1 + \dots) \right] \right\} \quad (4.25)$$

A computer for this equation may be realized as in Fig. 14.



p = PROBABILITY OF A CHANGE IN θ

$q = 1 - p = P(2)/P(1)$

FIG. 13. PROBABILITY COMPUTER, CASE 2b.

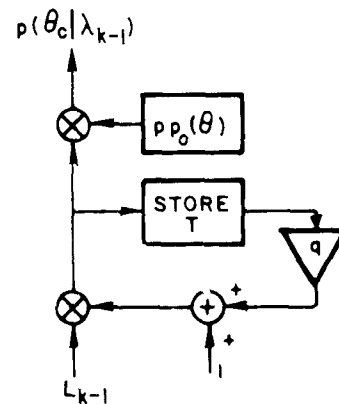
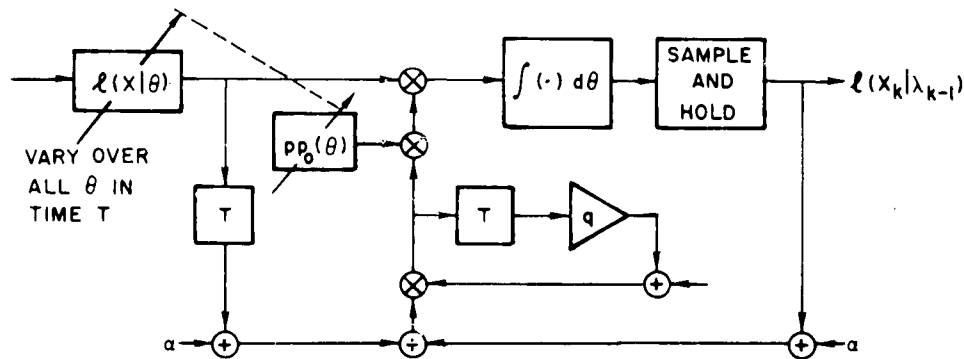


FIG. 14. PROBABILITY COMPUTER,
CASE 2c.

d. Independent Values--Scalar Parameters

When the unknown parameter is scalar (such as in the frequency-hopping problem), the system to compute $\ell(x_k | \lambda_{k-1})$ may be realized in the sweeping form shown in Fig. 15.



33367

FIG. 15. LEARNING SYSTEM, CASE 2c.

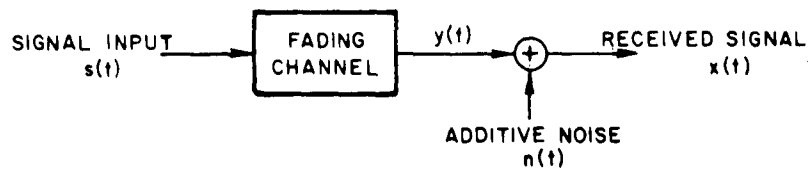
C. EXAMPLES

In order to demonstrate the utility of this synthesis technique when applied to problems in which the unknown parameter is time varying, consider the solution to the problems described briefly below.

1. The Fading-Channel Problem

In order to obtain a system which will be simple enough to illustrate the application of the foregoing technique, and at the same time realistic enough to demonstrate the utility of this technique, we shall utilize the following mathematical model of a data link using on-off keying for binary-coded transmission of data through a fading channel.[†] Figure 16 illustrates the channel.

[†]The channel model used is, according to Turin [Ref. 28], representative of propagation through the ionosphere above the MUF, or through the troposphere.



33366

FIG. 16. THE FADING-CHANNEL MODEL.

a. On-Off Keyed (OOK) Signals

(1) Signal. The information is transmitted as a sequence of marks and spaces. The signal is on for a duration T when a mark is being transmitted, and off for a duration T when a space is being transmitted. When the signal is on, it has the form

$$\hat{s}(t) = \text{Re} \{s(t) \exp(j\omega_0 t)\}$$

where ω_0 is known; $s(t)$ is a known, real, lowpass modulation waveform of duration T ; and Re denotes "real part of."

(2) Channel. The "nonselective, slow-fading" channel model used by Turin [Ref. 28][†] will be assumed. This channel is represented best by its operation on the signal. The channel output $y(t)$ may be represented as

$$y(t) = \text{Re} \{gs(t - \tau) \exp[j(\omega_0 t - \phi)]\}$$

(Thus by ignoring the modulation delay τ , we may think of the channel as a multiplicative medium with constant $G = ge^{-j\phi}$). The medium is characterized by the three quantities: g , the attenuation; τ , the modulation delay; and ϕ , the carrier phase shift. We assume that τ is known to the receiver, g is Rayleigh distributed and ϕ is uniformly distributed over the interval 0 to 2π . The channel is assumed to vary

[†]Turin discusses this model and the physical justification in considerable detail and therefore no attempt is made here to repeat his discussion.

slowly so that g and ϕ may be treated as constants over at least one signal duration T . More detailed time-variation assumptions will be made later as they are required.

The additive noise is assumed to be gaussian with constant spectral density $N_o/2$ over the narrow band of interest. Since $n(t)$ is a narrowband gaussian random process (NBGRP)[†] it may be written in terms of a complex modulation process as

$$n(t) = \text{Re} \{ \eta(t) \exp(j\omega_o t) \}$$

where $\eta(t)$ is a lowpass, complex, gaussian random process (GRP).

(3) Problem Formulation. The problem is to process the received waveform in a manner which will result in a minimum average risk decision. Because g is Rayleigh and ϕ is uniform, the quantity $gs(t)e^{-j\phi}$ is a lowpass complex GRP, and the quantity $gs(t)e^{-j\phi} + \eta(t)$ must be a lowpass GRP. We may note that $x(t)$ under either hypothesis may be written as the cisoid $\exp(j\omega_o t)$ modulated by a complex, low-pass GRP; hence $x(t)$ is an NBGRP.

If we utilize the complex notation

$$x(t) = \text{Re} \{ \zeta(t) \exp(j\omega_o t) \}$$

we may reformulate the hypothesis in terms of $\zeta(t)$ as follows:

$$H_1 \rightarrow \zeta(t) = gs(t) e^{-j\phi} + \eta(t)$$

$$H_2 \rightarrow \zeta(t) = \eta(t)$$

[†] A comprehensive discussion of the properties of narrowband gaussian random processes may be found in Refs. 25 and 26.

(We note, parenthetically, that we may obtain $\zeta(t)$ to a very good approximation from $x(t)$ utilizing the following equations:

$$\operatorname{Re} \{\zeta(t)\} \doteq \frac{2}{T} \int_{t-a}^t x(t) \cos \omega_0 t \, dt [= \check{x}(t)]$$

$$\operatorname{Im} \{\zeta(t)\} \doteq \frac{2}{T} \int_{t-a}^t x(t) \sin \omega_0 t \, dt [= \tilde{x}(t)]$$

where a is short compared to time variations in $s(t)$ and long compared to variations in $\cos \omega_0 t$. We denote these two real quantities by $\check{x}(t)$ and $\tilde{x}(t)$ respectively.)

It is shown in Ref. 26 that the real and imaginary parts of the lowpass complex envelope of an NBGRP are independent if they have symmetric spectral distribution; hence $\check{x}(t)$ and $\tilde{x}(t)$ are independent GRP's if we assume that the spectrum of the fading medium meets these requirements.

For brevity we denote by $(\check{})$ the real part and by $(\sim)$ the imaginary part. We note that $\tilde{s}(t)$ may be considered to be zero for on-off keyed signals, and denote

$$\check{g} = \operatorname{Re} g e^{-j\phi}$$

$$\tilde{g} = \operatorname{Im} g e^{-j\phi}$$

We may identify $\check{g}$ and $\tilde{g}$ as the in-phase and the quadrature channel gains. Then when H_1 is true,

$$\check{x}(t) = \check{g}\check{s}(t) + \check{\eta}(t)$$

$$\tilde{x}(t) = \tilde{g}\tilde{s}(t) + \tilde{\eta}(t)$$

We represent the k^{th} observation of $\check{x}(t)$ and $\tilde{x}(t)$ as the column vectors $\check{\mathbf{x}}_k$ and $\tilde{\mathbf{x}}_k$ which have as their rows the 2TW

samples of $\check{x}(t)$ and $\tilde{x}(t)$, $[(k-1)T \leq t < kT]$, sampled at the rate $2W$ samples per second [W is the bandwidth of the envelope $s(t)$]. Then the likelihood ratio, conditioned on $\check{g}_k$ and $\tilde{g}_k$, may be written as in Eq. (4.26) ($\check{g}_k$ and $\tilde{g}_k$ are the values of the unknown parameters $\check{g}$ and $\tilde{g}$ during the k^{th} observation).

$$\ell(\check{x}_k | \check{g}_k, \tilde{g}_k) = \ell(\check{X}_k, \tilde{X}_k | \check{g}_k, \tilde{g}_k) = \frac{p(\check{X}_k, \tilde{X}_k | \check{g}_k, \tilde{g}_k, H_1)}{p(\check{X}_k, \tilde{X}_k | \check{g}_k, \tilde{g}_k, H_2)} \quad (4.26)$$

But $\check{X}_k$ and $\tilde{X}_k$ are independent when H_2 is true, $\check{g}_k$ and $\tilde{g}_k$ are independent, $\check{X}_k$ does not depend on $\tilde{g}_k$, and $\tilde{X}_k$ does not depend on $\check{g}_k$; therefore we have

$$\begin{aligned} \ell(\check{X}_k, \tilde{X}_k | \check{g}_k, \tilde{g}_k) &= \frac{p(\check{X}_k | \check{g}_k, H_1) p(\tilde{X}_k | \tilde{g}_k, H_1)}{p(\check{X}_k | \check{g}_k, H_2) p(\tilde{X}_k | \tilde{g}_k, H_2)} \\ &= \ell(\check{X}_k | \check{g}_k) \ell(\tilde{X}_k | \tilde{g}_k) \end{aligned} \quad (4.27)$$

where, due to the normality of the noise,

$$\begin{aligned} \ell(\check{X}_k | \check{g}_k) &= \exp \left\{ \frac{-\check{g}_k^2}{2N_o W} S_t S + \frac{\check{g}_k}{N_o W} S_t \check{X}_k \right\} \\ \ell(\tilde{X}_k | \tilde{g}_k) &= \exp \left\{ \frac{-\tilde{g}_k^2}{2N_o W} S_t S + \frac{\tilde{g}_k}{N_o W} S_t \tilde{X}_k \right\} \\ \check{X}_k &= \begin{bmatrix} \check{x}_k(0) \\ \check{x}_k\left(\frac{1}{2W}\right) \\ \vdots \\ \check{x}_k\left(T - \frac{1}{2W}\right) \end{bmatrix}, \quad \tilde{X}_k = \begin{bmatrix} \tilde{x}_k(0) \\ \tilde{x}_k\left(\frac{1}{2W}\right) \\ \vdots \\ \tilde{x}_k\left(T - \frac{1}{2W}\right) \end{bmatrix}, \quad S = \begin{bmatrix} \hat{s}(0) \\ \hat{s}\left(\frac{1}{2W}\right) \\ \vdots \\ \hat{s}\left(T - \frac{1}{2W}\right) \end{bmatrix} \end{aligned}$$

The fact that the likelihood ratio factors into a part depending on $\check{\mathbf{g}}_k$ and a part depending on $\tilde{\mathbf{g}}_k$ will be useful in the synthesis of the system since it will allow synthesis of two independent systems, one to learn $\check{\mathbf{g}}_k$ and one to learn $\tilde{\mathbf{g}}_k$.

The optimum system computes

$$\ell(\mathbf{x}_k | \lambda_{k-1}) = \iint p(\check{\mathbf{g}}_k, \tilde{\mathbf{g}}_k | \lambda_{k-1}) \ell(\check{\mathbf{x}}_k | \check{\mathbf{g}}_k) \ell(\tilde{\mathbf{x}}_k | \tilde{\mathbf{g}}_k) d\check{\mathbf{g}}_k d\tilde{\mathbf{g}}_k \quad (4.28)$$

thus we require $p(\check{\mathbf{g}}_k, \tilde{\mathbf{g}}_k | \lambda_{k-1})$ in order to synthesize the system. It may be shown that $\check{\mathbf{g}}_k$ and $\tilde{\mathbf{g}}_k$ are conditionally independent:[†]

$$p(\check{\mathbf{g}}_k, \tilde{\mathbf{g}}_k | \lambda_{k-1}) = p(\check{\mathbf{g}}_k | \check{\lambda}_{k-1}) p(\tilde{\mathbf{g}}_k | \tilde{\lambda}_{k-1}) \quad (4.29)$$

where $\check{\lambda}_{k-1} = (\check{x}_1, \check{x}_2, \dots, \check{x}_{k-1})$

$$\tilde{\lambda}_{k-1} = (\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_{k-1})$$

Hence the system may be synthesized in the form shown in Fig. 17.

[†]The left-hand element of Eq. (4.29) can be written as

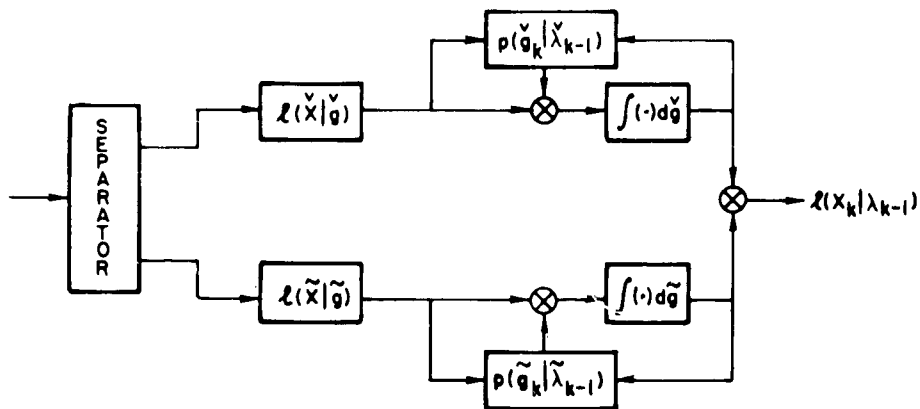
$$p(\check{\mathbf{g}}_k, \tilde{\mathbf{g}}_k | \lambda_{k-1}) = p(\check{\mathbf{g}}_k | \tilde{\mathbf{g}}_k, \lambda_{k-1}) p(\tilde{\mathbf{g}}_k | \lambda_{k-1})$$

Since knowing the value of λ_{k-1} is the same as knowing the value of $\check{\lambda}_{k-1}$ and $\tilde{\lambda}_{k-1}$, we may replace $p(\check{\mathbf{g}}_k | \tilde{\mathbf{g}}_k, \lambda_{k-1})$ by $p(\check{\mathbf{g}}_k | \tilde{\mathbf{g}}_k, \check{\lambda}_{k-1}, \tilde{\lambda}_{k-1})$. Alternatively,

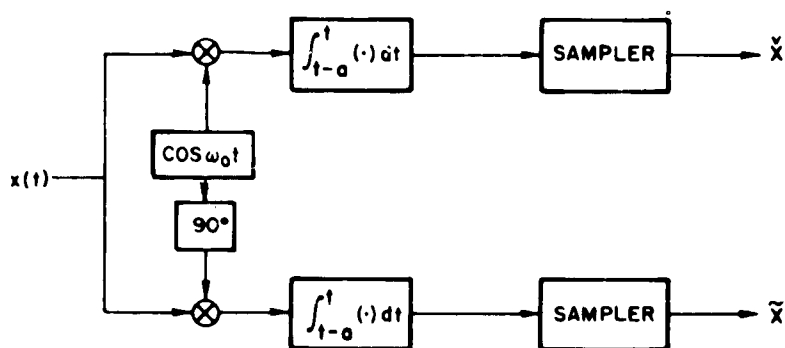
$$p(\check{\mathbf{g}}_k | \tilde{\mathbf{g}}_k, \check{\lambda}_{k-1}, \tilde{\lambda}_{k-1}) = \frac{p(\check{\mathbf{g}}_k, \check{\lambda}_{k-1} | \tilde{\mathbf{g}}_k, \tilde{\lambda}_{k-1})}{p(\check{\lambda}_{k-1} | \tilde{\mathbf{g}}_k, \tilde{\lambda}_{k-1})}$$

Since both $\check{\mathbf{g}}_k$ and $\check{\lambda}_{k-1}$ are independent of $\tilde{\mathbf{g}}_k$ and $\tilde{\lambda}_{k-1}$, then

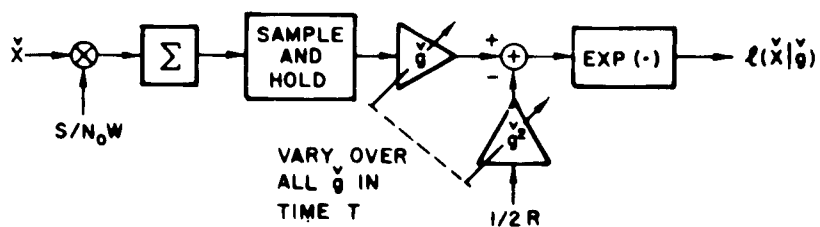
$$p(\check{\mathbf{g}}_k | \tilde{\mathbf{g}}_k, \lambda_{k-1}) = p(\check{\mathbf{g}}_k | \check{\lambda}_{k-1}) \quad \text{and} \quad p(\tilde{\mathbf{g}}_k | \lambda_{k-1}) = p(\tilde{\mathbf{g}}_k | \tilde{\lambda}_{k-1})$$



a. Fading on-off keyed signals



b. Separator



c. Conditional likelihood computer

33369

FIG. 17. LEARNING RECEIVER FOR FADING ON-OFF KEYED SIGNALS.

To determine the block diagram form of the box which computes $p(\check{g}_k | \check{\lambda}_{k-1})$, a model of the time variations of $\check{g}_k$ and $\tilde{g}_k$ is required. Two possibilities will be considered.

Case 1: The first and simpler of the two models involves the assumption that the fading process is a random-walk process; that is, assume that changes in $\check{g}$ and $\tilde{g}$ take place slowly enough so that each new value of either $\check{g}$ or $\tilde{g}$ is a small independent perturbation of the preceding value, so that

$$\check{g}_k = \check{g}_{k-1} + \check{\epsilon}_k \quad (4.30)$$

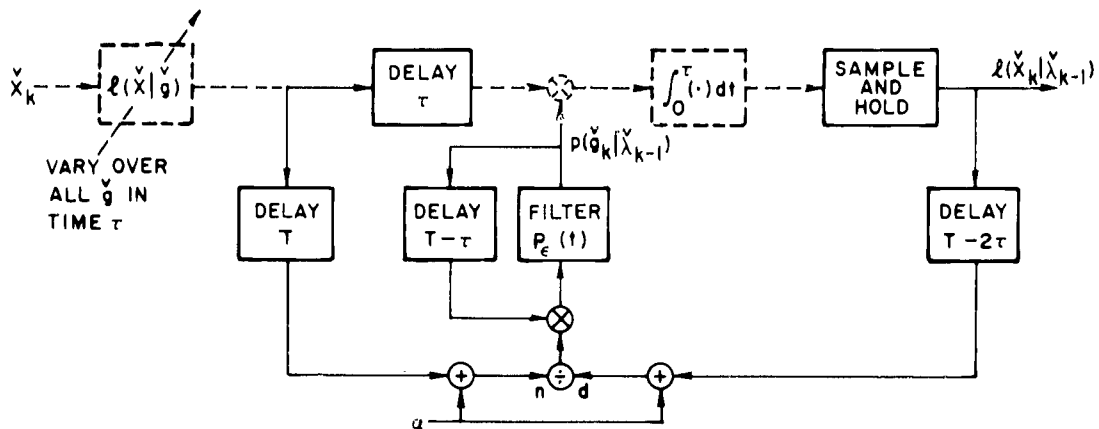
$$\tilde{g}_k = \tilde{g}_{k-1} + \tilde{\epsilon}_k$$

where $\check{\epsilon}_k$ is independent of $\check{g}_{k-1}$; $\tilde{\epsilon}_k$ is independent of $\tilde{g}_{k-1}$; and both are distributed according to $p_{\check{\epsilon}}(z) = p_{\tilde{\epsilon}}(z)$.

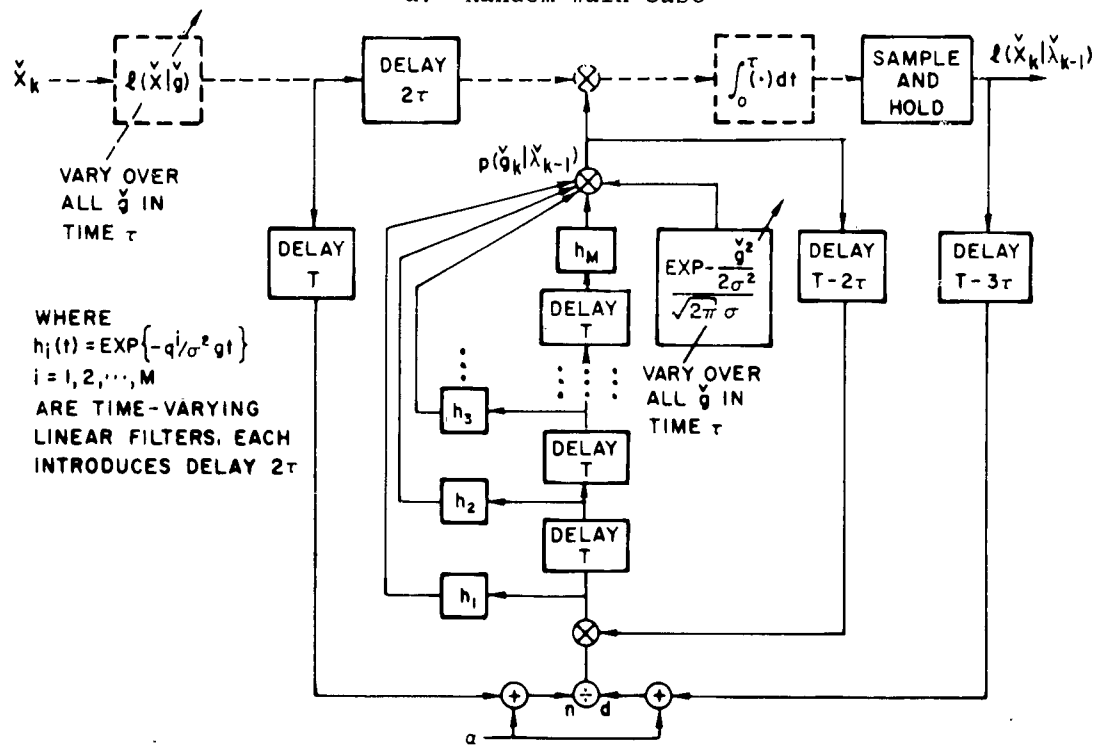
In this case, the box to compute $p(\check{g}_k | \check{\lambda}_{k-1})$ or $p(\tilde{g}_k | \tilde{\lambda}_{k-1})$ may be realized as shown in Fig. 18a.

Case 2: The second model is more involved, and allows the correlation between present and past values of $\check{g}$ and $\tilde{g}$ to be taken into account by treating the processes as M^{th} -order Markov variables. To be specific, assume that the correlation of $\check{g}$ or $\tilde{g}$ decreases exponentially with time back MT sec, and then becomes zero. In this case

$$p(\check{g}_k | \check{g}_{k-1}, \dots, \check{g}_{k-M}) = \frac{1}{\sqrt{2\pi\sigma}} \exp \left\{ -\frac{1}{2\sigma^2} \left[\check{g}_k^2 + 2\check{g}_k\check{g}_{k-1}^q + \dots + 2\check{g}_k\check{g}_{k-1}^q + \dots + 2\check{g}_k\check{g}_{k-M}^q \right] \right\} \quad (4.31)$$



a. Random-walk case



b. M^{th} -order Markov case

33370

FIG. 18. PROBABILITY COMPUTER.

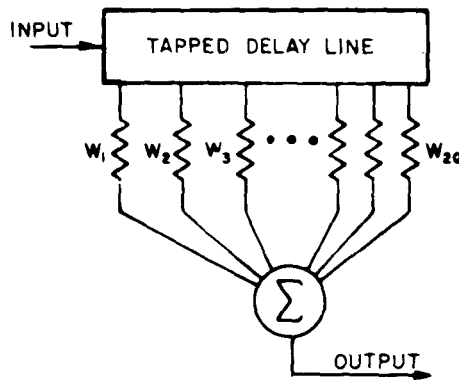
so that [see Eq. (4.15)]

$$\begin{aligned}
 p(\check{g}_k | \check{\lambda}_{k-1}) &= \int \dots \int p(\check{g}_k | \check{g}_{k-1}, \dots, \check{g}_{k-M}) \prod_{i=1}^M p(\check{g}_{k-i} | \check{\lambda}_{k-1}) d\check{g}_{k-1} \dots d\check{g}_{k-M} \\
 &= \frac{1}{\sqrt{2\pi\sigma}} \exp\left(\frac{-\check{g}_k^2}{2\sigma^2}\right) \prod_{i=1}^M \int p(\check{g}_{k-i} | \check{\lambda}_{k-1}) \exp\left(\frac{-q^i}{\sigma^2} \check{g}_k \check{g}_{k-i}\right) d\check{g}_{k-i}
 \end{aligned} \tag{4.32}$$

The system shown in Fig. 18b will compute this function. In this case it is necessary to utilize M time-varying, linear filters $h_1, \dots, h_M$. These filters have an impulse response

$$h_1(t, \check{g}_k) = \exp\left(\frac{-q^1}{\sigma^2} t \check{g}_k\right) \tag{4.33}$$

Such time-varying filters can be realized with a tapped delay line of delay length 2τ (where τ is the time required to sweep through the range of $\check{g}$). If the range of $\check{g}$ is quantized into Q levels, the filter will require $2Q$ taps, as shown in Fig. 19.



$$\begin{aligned}
 w_n &= \exp [-(nq^1/\sigma^2)\delta] \\
 n &= 1, 2, \dots, 2Q \\
 \delta &= \text{INCREMENTAL DELAY} = \tau/Q
 \end{aligned}$$

33382

FIG. 19. TAPPED-DELAY-LINE REALIZATION OF TIME-VARYING LINEAR FILTER.

b. Frequency-Shift Keyed (FSK) Signals

A somewhat more complicated, and perhaps more useful, example results when the signal model is modified so that the modulation is frequency-shift keying instead of on-off keying. Such a signal model is described below.

(1) Signal. The signal is on continuously; however it is shifted between two frequencies depending upon whether a mark or a space is being transmitted. This shift occurs at multiples of T . During transmission of a mark signal, $s_1(t)$ is transmitted; and during a space, $s_2(t)$ is transmitted, where:

$$s_1(t) = \operatorname{Re} \{ S \exp (j\omega_1 t) \} \quad 0 \leq t \leq T$$

$$s_2(t) = \operatorname{Re} \{ S \exp (j\omega_2 t) \} \quad 0 \leq t \leq T$$

We assume that ω_1 and ω_2 are chosen so that the signals are orthogonal over the interval T ; i.e.,

$$\int_0^T s_1(t) s_2(t) dt = 0 \quad (4.34)$$

(2) Channel. We make the same assumptions concerning the channel as for the on-off keyed signal. In this case, however, there are two channels of interest, one at ω_1 and the other at ω_2 . We assume that the two channels fade independently and that the multiplicative constants $G_1 = g_1 \exp (-j\phi_1)$ and $G_2 = g_2 \exp (-j\phi_2)$ are independent, complex gaussian random processes with symmetric spectral distributions.

(3) Problem Formulation. In this case there are two hypotheses:

$$H_1 = \text{hypothesis that } x(t) = G_1 s_1(t) + n(t)$$

$$H_2 = \text{hypothesis that } x(t) = G_2 s_2(t) + n(t)$$

By writing $x(t)$ in the complex modulation form at the two frequencies and taking advantage of (1) the narrowband nature of the processes and (2) the independence of the channels, it is readily shown that the likelihood ratio factors. Similarly, the joint conditional probability density $p(\check{g}_{1,k}, \check{g}_{1,k}, \check{g}_{2,k}, \check{g}_{2,k} | \lambda_{k-1})$ factors, so that

$$\ell(x_k | \lambda_{k-1}) = [\ell(\check{X}_{1,k} | \check{\lambda}_{1,k-1}) \ell(\tilde{X}_{1,k} | \tilde{\lambda}_{1,k-1})] [\ell(\check{X}_{2,k} | \check{\lambda}_{2,k-1}) \ell(\tilde{X}_{2,k} | \tilde{\lambda}_{2,k-1})] \quad (4.35)$$

where

$$\ell(\check{X}_{1,k} | \check{\lambda}_{1,k-1}) = \ell(\check{X}_{1,k} | \check{g}_{1,k}) p(\check{g}_{1,k} | \check{\lambda}_{1,k-1}) d\check{g}_{1,k}$$

$$\check{X}_{1,k} = \begin{bmatrix} \check{x}_i(0) \\ \check{x}_i(\Delta) \\ \vdots \\ \check{x}_i(T - \Delta) \end{bmatrix} ; \quad \tilde{X}_{1,k} = \begin{bmatrix} \tilde{x}_i(0) \\ \tilde{x}_i(\Delta) \\ \vdots \\ \tilde{x}_i(T - \Delta) \end{bmatrix}$$

$$\check{x}_i(t) = \int_{t-a}^t x(t) \cos \omega_i t dt$$

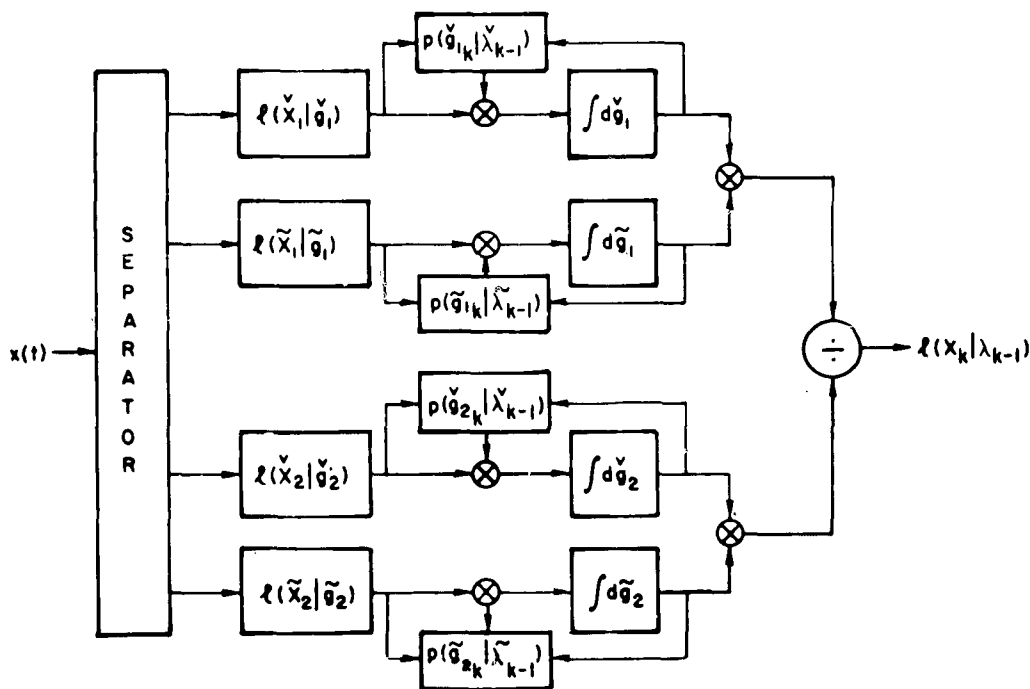
$$\tilde{x}_i(t) = \int_{t-a}^t x(t) \sin \omega_i t dt$$

$$\check{\lambda}_{i,k} = (\check{X}_{i,1}, \check{X}_{i,2}, \dots, \check{X}_{i,k})$$

$$\tilde{\lambda}_{i,k} = (\tilde{X}_{i,1}, \tilde{X}_{i,2}, \dots, \tilde{X}_{i,k})$$

and where $i = 1, 2$.

Thus the solution to the independent fading-channel problem when FSK modulation is used is the ratio of the output of two of the previous systems. (See Fig. 20.)



33371

FIG. 20. LEARNING RECEIVER FOR FADING FREQUENCY-SHIFT KEYED SIGNALS.

2. Frequency-Hopping Signal Reconnaissance Problem

There are many reconnaissance problems in which it is desired to detect the presence of a signal with unknown or randomly time-varying parameters. Such problems are often readily solved by the procedures outlined above. One such problem involves the detection of a frequency-hopping signal embedded in noise. The model for this example follows.

a. Signal

The signal is assumed to be a narrowband signal which may be represented over an interval of duration T by a sample function of a narrowband gaussian random process with center frequency ω which is an unknown, time-varying parameter. Hence

$$S(t) = a \cos (\omega t + \phi)$$

where a is Rayleigh distributed, $p(a) = (a/A^2) \exp (-a^2/2A^2)$, and ϕ is uniformly distributed over the interval from 0 to 2π . The frequency is assumed to change only at integral multiples of the interval T . The probability of a change in frequency is $p \ll 1$ independently of when the last change occurred, and the frequency is equally likely to change to any value within a specified band W .

b. Noise

The noise is normally distributed, with constant spectral density $N_o/2$ over the band W .

c. Problem Formulation

The problem is to examine intervals (of duration T) of the received waveform and to make a signal-presence decision at the end of each interval; hence signal-present and signal-absent hypotheses are defined as in example 1. Because the unknown variable is a scalar with zero-order Markov value dependence and binomial time dependence, the system for detection must take the form of the system of Fig. 15, with θ replaced by f . To complete the solution, an expression for $\ell(X|f)$ is required. This expression is (see Chapter II)

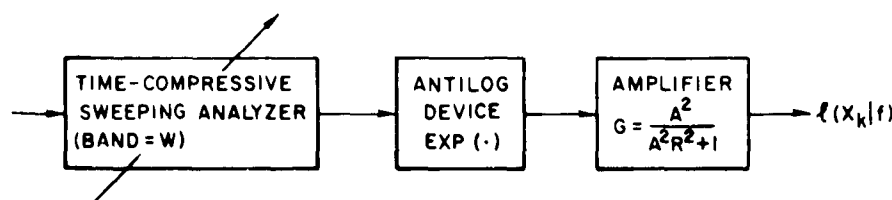
$$\ell(X|f) = \frac{1}{\frac{A^2}{2N_o W} + 1} \exp \left[\frac{\frac{A^2}{\frac{A^2}{2N_o W} + 1} |X_t E(f)|^2 \right] \quad (4.36)$$

where

$$X = \begin{bmatrix} x(0) \\ x(\Delta) \\ \vdots \\ x(T - \Delta) \end{bmatrix}, \quad E(f) = \begin{bmatrix} 1 \\ \exp(j2\pi f \Delta) \\ \vdots \\ \exp[j2\pi f(T - \Delta)] \end{bmatrix}$$

$x(t)$ = received signal, $\Delta = \frac{1}{2W}$ = sampling interval

The quantity $|X_t E(f)|^2$ is proportional to the periodogram of the input at frequency f , which in turn is closely related to the spectral density of $x(t)$ at f .[†] From these facts it may be shown that the likelihood computer (which must sweep over the range of f) consists of a time-compressive sweeping spectrum analyzer[†] followed by an antilog device and an amplifier, as shown in Fig. 21. Here the sweeping analyzer must cover the band W in the time T , and repeat periodically.



33372

FIG. 21. LIKELIHOOD COMPUTER.

A receiver of this nature will optimally detect frequency-hopping signals for which the model proposed is a suitable representation. Although it is more complicated than many receivers, such an adaptive receiver should not be particularly difficult to construct.

D. SUMMARY OF CHAPTER IV

In this chapter we have investigated the learning problem in which the unknown parameter is time varying. By utilizing two specific models for the way in which the parameter may vary in time, we have demonstrated that the same techniques which are applicable to the solution of learning problems when the parameter is fixed are applicable when the parameter varies in time. Furthermore, through the use of two examples, we have demonstrated that the models proposed are applicable in a variety of physical situations.

[†] For a general discussion of the periodogram see Ref. 25; for a discussion of a time-compressive spectrum analyzer see Ref. 29; and for the relationship between spectral analysis and the periodogram see Ref. 30.

V. SYSTEM REALIZABILITY

The purpose of this chapter is to investigate the physical realizability of the optimum learning systems developed in the previous chapters. The realizability of a system will be defined in terms of the number of elements required to construct the system rather than in terms of the realizability of the individual elements. A system which requires a finite number of perfect elements such as amplifiers, multipliers, adders, storage elements, etc., will be considered to be realizable.

It is important to recall that very few, if any, mathematical models are exact representations of a physical problem, although the models may be accurate enough that the difference between physical and theoretically predicted events cannot be measured. Such models are considered to be adequate representations in an "engineering" sense. It is in this engineering sense that the individual elements of the learning systems are physically realizable, and it is in this sense that we shall demonstrate the realizability of many learning systems.

A. SYSTEM MEMORY CAPACITY

Learning systems extract and store information from a sequence of observations. They are useful if the information storage required is less than the storage required to store the observation sequence. In the systems developed in Chapters II and IV the system size (number of elements) depends directly on the number of information-storage elements required. From the mathematical description of the systems, it is clear that the information stored is used to compute $p(\theta|\lambda_k)$; thus to investigate system size we investigate the memory capacity M_c required to compute $p(\theta|\lambda_k)$. We define the required M_c of an optimal learning machine as the minimum number of functions $\phi_i(\lambda_k)$ of the observation sequence λ_k which must be stored by the learning machine.

In order to investigate the theoretical information-storage capacity required, we shall examine the concept of necessary and sufficient (minimal sufficient) statistics, and the dimensionality of the linear space spanned by these statistics. We shall utilize the definitions of

Dynkin [Ref. 31] and Grettenberg [Ref. 32] to prove the following statements:

1. The system which computes $p(\theta|\lambda_k)$ computes a minimal sufficient statistic.
2. No optimal learning system may be constructed with a memory capacity less than the memory capacity required to compute $p(\theta|\lambda_k)$.
3. If the set Φ of all possible values of the unknown parameter θ consists of Q points $\theta_1, \theta_2, \dots, \theta_Q$, the memory capacity of an optimal learning machine is less than or equal to $Q-1$.

We shall show in Sec. C that in many learning problems a discrete model for Φ exists which is adequate in an engineering sense.

B. MINIMAL SUFFICIENT STATISTICS

Systems to solve the classification problem when an important parameter is unknown must extract and store certain information from a sequence of observations. The information to be stored is that which will allow the selection of the conditional probability distribution $p(X|\theta)$ (from which the observation X was drawn) from a family of distributions indexed by θ . Systems which perform this selection are computing functions of the learning observations which partition the observation space into a set of decision regions. It is well known that certain functions of the learning sequence lead to Bayes decision regions regardless of the loss functions and a priori probabilities [Ref. 33]. Such functions are sufficient to make a minimum average risk decision, hence they are called sufficient statistics.

Some sufficient statistics are more desirable to compute than others because they require the storage of less information. Since the learning problem under study requires a sufficient statistic, it is desirable to choose that one which requires the least information storage. A function within this class is called by Dynkin [Ref. 31] a necessary and sufficient statistic; however a more descriptive name, which has been used by Grettenberg [Ref. 32], is a "minimal sufficient statistic."

A sufficient statistic, in the above sense, may be defined[†] as follows.

Definition: A statistic $T(X)$ is sufficient for the family $\{p(X|\theta) : \theta \in \Phi\}$ if and only if $p(X|\theta)$ may be factored as follows

$$p(X|\theta) = h(X) f(T(X), \theta) \quad (5.1)$$

where $h(X)$ depends only on X and $f(T, \theta)$ depends on X only through T .

In order to study minimal sufficient statistics, we first define these functions in terms of functional dependence as below.

Definition: A sufficient statistic $T_1(X)$ is dependent on another sufficient statistic $T_2(X)$ if $T_2(X_1) = T_2(X_2)$ implies $T_1(X_1) = T_1(X_2)$, that is, if $T_1(X)$ may be written as a function which depends on X only through $T_2(X)$.

Definition: A minimal sufficient statistic $T(X)$ is a sufficient statistic which depends on all other sufficient statistics.

From these definitions it is clear that the function $p(\theta|\lambda_k)$ is a minimal sufficient statistic for the family $\{p(X|\theta) : \theta \in \Phi\}$, and a sample of size k . That is, it is sufficient because $p(X_1, \dots, X_k|\theta)$ may be factored as

$$p(X_1, \dots, X_k|\theta) = \frac{p(\theta|\lambda_k)}{p_o(\theta)} p(X_1, \dots, X_k) \quad (5.2)$$

and it is minimal since it depends on every other sufficient statistic. To show this, let $T(X_1, \dots, X_k)$ be a sufficient statistic, then

$$p(X_1, \dots, X_k|\theta) = h(X_1, \dots, X_k) f(T(X_1, \dots, X_k), \theta) \quad (5.3)$$

[†]The concept of sufficient statistics is only interesting when we observe more than one sample. We may define minimal sufficient statistics for the sample of size n by replacing $T(X)$ by $T(X_1, X_2, \dots, X_n)$ in each of the definitions given.

so that

$$\begin{aligned}
 p(\theta | X_1, \dots, X_k) &= \frac{p(X_1, \dots, X_k | \theta) p_o(\theta)}{\int p(X_1, \dots, X_k | \theta) p_o(\theta) d\theta} \\
 &= \frac{h(X_1, \dots, X_k) f(T(X_1, \dots, X_k), \theta) p_o(\theta)}{h(X_1, \dots, X_k) \int f(T(X_1, \dots, X_k), \theta) p_o(\theta) d\theta}
 \end{aligned}
 \tag{5.4}$$

Hence $p(\theta | \lambda_k)$ depends on $\{X_1, \dots, X_k\}$ only through $T(X_1, \dots, X_k)$.

Thus the optimal learning system computes a minimal sufficient statistic, and the first statement of Sec. A has been demonstrated. To demonstrate the second statement we proceed as follows.

An optimal learning machine for the observation sequence λ_k and the family $\{p(X|\theta) : \theta \in \Phi\}$ must compute a sufficient statistic of λ_k . A minimal sufficient statistic of λ_k is a many-one transformation on all other sufficient statistics (except other minimal sufficient statistics) because it is functionally dependent on all other sufficient statistics. Thus the number of functionally independent functions of λ_k which must be computed to compute a minimal sufficient statistic must be minimal. No optimal learning machine can be constructed with a memory capacity less than that of a machine which computes a minimal sufficient statistic.

Finally, we shall demonstrate that the memory capacity M_c of a machine to compute $p(\theta | \lambda_k)$ is finite whenever the set Φ of all possible values of θ consists of Q points $\theta_1, \theta_2, \dots, \theta_Q$, and that $M_c \leq Q-1$. The function $p(\theta | \lambda_k)$ may be written

$$p(\theta | \lambda_k) = \sum_{i=1}^Q I_i(\theta) g_i(\lambda_k)
 \tag{5.5}$$

where

$$I_i(\theta) = \begin{cases} 1 & \theta = \theta_i \\ 0 & \text{elsewhere} \end{cases}$$

$$g_i(\lambda_k) = P(\theta_i | \lambda_k)$$

But

$$\begin{aligned} P(\theta_i | \lambda_k) &= \prod_{j=1}^k p(X_j | \theta_i) P_o(\theta_i) \\ &= \exp \left\{ \sum_{j=1}^k \ln p(X_j | \theta_i) \right\} P_o(\theta_i) \end{aligned} \quad (5.6)$$

Thus it is sufficient to store the Q functions

$$\sum_{j=1}^k \ln p(X_j | \theta_i)$$

in order to be able to compute $p(\theta | \lambda_k)$ and $M_c \leq Q-1$.

If we have the case where the functions $p(X | \theta_i)$ are functionally independent, then it is necessary to store the Q functions, and the inequality ($M_c \leq Q-1$) becomes an equality ($M_c = Q-1$).

It is clear that once we are given a decision problem involving $\{p(X | \theta) : \theta \in \Phi\}$, we may readily construct a finite-sized system so long as Φ is a set of Q points. In fact, by taking advantage of any functional dependence which may exist between the functions $p(X | \theta_i)$, we may always construct a system which requires a minimum of information storage capacity.

The problems in which Φ does not consist of a discrete set of points may often lead to systems in which M_c is not finite. Thus the systems will not be realizable.

C. PRACTICAL CONSIDERATIONS

We have just noted that the memory size of the learning system is finite so long as the unknown parameter space is discrete, and that in many cases of interest it is infinite when the parameter space is not discrete. Since it is not usually considered possible to construct systems with infinite memory capacity, we may draw the conclusion that we cannot construct the theoretically optimum system in these cases and can then set about either changing the theoretical model, or looking for a suboptimum finite system.

One reasonable way in which to modify the model is to ask for the optimum (Bayes) system under a finite memory constraint. Such an approach, although logical, is difficult to apply to the learning problem and will not be attempted in this study.

Instead of attempting to modify the model we may find it more useful to examine the results of simply using the model to synthesize optimum systems, and then to approximate these systems as well as we can. Although this approach is much less pleasing mathematically, it has the advantage of being practical, and has some precedent in other applied decision-theory fields.

A similar situation exists whenever we represent a continuous function $x(t)$ in the interval $(0, T)$ by its sample values $x(0), x(t_1), \dots, x(t_m)$ taken in this interval. In an engineering sense for some large m these samples adequately specify $x(t) : t \in (0, T)$; however, strictly speaking, unless $m \rightarrow \infty$ this is only an approximation [see Ref. 20].

The fact that in most cases there is some finite set of discrete values of the unknown parameter which in an engineering sense represent all of the usefully distinguishable values that the parameter may assume is stated in the following theorem. A system based on this set of possible parameter values requires a finite (fixed) memory capacity.

Theorem 1. Designate by Φ the space of all possible values of the vector parameter θ , and let the range of each coordinate of θ be bounded. Then if $p(X|\theta, H_2)$ is independent of θ , and $p(X|\theta, H_1)$ is a continuous function of θ for all $\theta \in \Phi$ and all X , there exists a subset of Φ , say $\Phi_Q = \{\theta_1, \theta_2, \dots, \theta_Q\}$, with a finite number, Q , of discrete values of θ such that for any $\epsilon > 0$ and all $\hat{\theta} \in \Phi$ there is a $\theta_q \in \Phi_Q$ which satisfies

$$\rho(d^*(\theta_q)|\hat{\theta}) \leq \rho(d^*(\hat{\theta})|\hat{\theta}) + \epsilon$$

where $\rho(d^*(\theta_q)|\hat{\theta})$ is the average risk of the Bayes decision rule based on the assumption that θ_q is true ($d^*(\theta_q)$) when $\hat{\theta}$ is true.

This theorem is proven in Appendix B. The condition that $p(X|\theta, H_1)$ be a continuous function of θ is not particularly restrictive and could be removed by first extracting the set of θ at which any discontinuities exist, provided this set is finite. The condition that $p(X|\theta, H_2)$ be independent of θ has been introduced primarily to simplify the proof of the theorem. In most applications this condition will be met. If it is not, it can be replaced by the requirement that $p(X|\theta, H_2)$ be a continuous function of θ . These conditions are all usually met in applications so that a finite set Φ_Q will almost always exist in practice.

This theorem demonstrates that in many binary decision problems we may quantize the space of the unknown parameter in such a way that if a learning system is constructed on the basis of this quantization, and if the learning system converges so that it utilizes $d^*(\theta_q)$, the ultimate system performance will be arbitrarily close to the ultimate system performance of a system based on the unquantized space. In Chapter VI we shall demonstrate that in most binary decision problems the "quantized" system will converge to $d^*(\theta_q)$ such that

$$\rho(d^*(\theta_q)|\hat{\theta}) = \min_{\theta_q \in \Phi_Q} \rho(d^*(\theta_q)|\hat{\theta})$$

In order to illustrate the choice of quantization coarseness, consider the following example.

Example: Suppose that we wish to detect an unknown signal in gaussian noise. Let $\hat{S}$ = signal vector

K = noise covariance matrix

S = unknown parameter

Then we know [see Ref. 3 or 23] that the quality of performance of a system which is given a priori knowledge of $\hat{S}$ is dependent on the "divergence" defined by

$$d^2 = \hat{S}_t K^{-1} \hat{S}$$

The quality of performance of any other linear system using a slightly mismatched filter can be measured by the ratio of divergences.

The difference in performance of two systems is a continuous monotonic function of this ratio, and is zero when the ratio is 1. If we require that

$$\left(\frac{d'}{d}\right)^2 > 1 - \epsilon \quad (5.7)$$

where ϵ is a small quantity, the performance difference will be small. Thus if we have a system in which the S -space is quantized so that the nearest point to $\hat{S}$ is, say, $S^* = \hat{S} + \Delta$, then [Ref. 3, p. 45]

$$\begin{aligned} \left(\frac{d'}{d}\right)^2 &= \frac{(\hat{S}_t K^{-1} S^*)^2}{(\hat{S}_t K^{-1} \hat{S})(S^{*t} K^{-1} S^*)} \\ &= \frac{(\hat{S}_t K^{-1} \hat{S})^2 + 2(\hat{S}_t K^{-1} \hat{S})(\hat{S}_t K^{-1} \Delta) + (\hat{S}_t K^{-1} \Delta)^2}{(\hat{S}_t K^{-1} \hat{S})(\hat{S}_t K^{-1} \hat{S} + 2\hat{S}_t K^{-1} \Delta + \Delta_t K^{-1} \Delta)} \end{aligned} \quad (5.8)$$

Relations (5.7) and (5.8) require that

$$\left(\hat{S}_t K^{-1} \hat{S}\right) \left(S_t^* K^{-1} S^*\right) - \left(\hat{S}_t K^{-1} S^*\right) \left(S_t^* K^{-1} \hat{S}\right) < \epsilon \left(\hat{S}_t K^{-1} \hat{S}\right) \left(S_t^* K^{-1} S^*\right) \quad (5.9a)$$

This relation may be simplified, if $\left(\hat{S}_t K^{-1} \hat{S}\right) \left(S_t^* K^{-1} S^*\right)$ is greater than one, to yield

$$\left(\hat{S}_t K^{-1} \hat{S}\right) \left(\Delta K^{-1} \Delta\right) - \left(\hat{S} K^{-1} \Delta\right) < \epsilon' \quad (5.9b)$$

We may use any quantization interval in S-space which satisfies (5.9a) or (5.9b) as appropriate. The resulting system will be capable of performing nearly as well in the steady state as a system with a priori knowledge of $\hat{S}$.

D. SUMMARY OF CHAPTER V

In this chapter we have discussed the question of system realizability in terms of the number of information storage elements required of an optimal learning system. We have been able to prove two important facts about learning problems.

1. Learning systems to solve problems in which the unknown parameter may take on only a finite number of values are always finite in size.
2. Most learning problems in which the unknown parameter may take on an infinite number of values may be adequately represented by problems in which the number of values is finite.

In the second statement an adequate representation is one which leads to a system which will perform almost as well as the system based on the infinite model. The second statement depends upon the fact that the system based on the finite model will converge even though the infinite model is the best representation of the physical problem.

In the second statement the existence of an adequate representation means that for every possible value of the unknown parameter in the infinite set there is a value in a finite set which is arbitrarily "close"

when "distance" is measured in terms of the difference in performance of the corresponding systems. This latter statement will become particularly meaningful in the next chapter when we show that learning systems based on the finite-set representation will converge to the finite system which is "closest" in a performance sense to the optimum system based on the infinite set.

VI. SOME PROPERTIES OF LEARNING SYSTEMS

Systems which have been synthesized as proposed in Chapters II and IV have several interesting properties. Such systems are stable, and they converge to the system which would be optimum if the unknown parameter were known. Furthermore, systems which are constructed as suggested in Chapter V, by quantization of the unknown parameter, also converge to the discrete point in the quantized space which is nearest the convergence point of the equivalent nonquantized system. A most interesting property of the recursive expressions developed in Chapters II and IV is the fact that in addition to being applicable to the problems of those chapters they are also generally applicable to problems in which learning with a teacher is possible and to problems in which no learning is possible.

It is the purpose of this chapter to formalize the statement of these properties, and to specify the conditions under which they hold. For convenience, we shall carry out the following discussion in terms of the binary decision problem since with a few obvious changes the discussion would apply equally well to the more general solution.

A. SYSTEM STABILITY

Because the system requires both delay feedback and feedforward loops, the question arises whether or not there is an input sequence which can cause an output which will be unbounded. Although we cannot answer this stability question in the normal control-system manner, we can provide a satisfactory answer in probabilistic terms; that is, we can show that the probability that the output will grow without bound is zero. We can obtain this answer by showing that the sequence of outputs $E(X_k | \lambda_{k-1})$ is a bounded martingale (Appendix C) when $E(X|\theta)$ is bounded for all θ and fixed X . Since bounded martingales have the property that they are bounded for all sequences $\{X_k, \lambda_{k-1}\}$ with probability one [Ref. 31], we will have answered the stability question if we can show that $E(X|\theta)$ is bounded. But certainly this must be true unless the signal is "perfectly detectable," and this is a pathological situation which seems to occur only in textbooks.

As an example of the boundedness of $\ell(X|\theta)$ consider example 1 of Chapter II. In this case we identify θ with the unknown amplitude, c .

$$\ell(X|c) = \exp \left(-\frac{1}{2} c^2 B_t K^{-1} B + c X_t K^{-1} B \right) \quad r_1 \leq c \leq r_2 \quad (6.1)$$

Given any input vector X_0 , this function is certainly bounded by

$$\exp \left(-\frac{1}{2} r_1^2 B_t K^{-1} B + r_2 |X_0 K^{-1} B| \right) \quad (6.2)$$

for all values of c .

B. CONVERGENCE OF THE CONTINUOUS SYSTEM

In Chapter II we described systems for the solution of problems in which an important parameter was fixed but unknown. An important property of such systems is the fact that they converge, so that in a sense they "learn" the fixed value of the parameter. In Chapter IV we described similar systems for similar problems in which the difference was the fact that the parameter was time varying. Since the parameter varies with time, we cannot discuss the steady-state performance of these systems, and therefore the following discussion is applicable only to the systems of Chapter II.

We investigate the convergence by again appealing to the martingale nature of the output. In Appendix C we show that if a sequence of functions $\{\phi_k(X_1, \dots, X_k)\}$ exists such that

$$\lim_{k \rightarrow \infty} \phi_k(X_1, \dots, X_k) = \hat{\theta} \quad \text{with probability one} \quad (6.3)$$

then

$$\lim_{k \rightarrow \infty} i(X_k | \lambda_{k-1}) = i(X | \hat{\theta}) \quad \text{with probability one} \quad (6.4)$$

where $\hat{\theta}$ is the true value of θ . Thus the system (in the limit) performs as well as one which was designed with knowledge of the signal.

As an example of a problem in which the sequence $\{\phi_k\}$ exists, we may consider the problem of detection of an unknown signal in noise. Consider the linear estimate of S (the signal) given by

$$\hat{S}_k = \frac{1}{kp(H_1)} \sum_{i=1}^k X_i \quad (6.5)$$

The observations X_i may be written as

$$X_i = N_i + Y_i S_i \quad (6.6)$$

where

$$Y_i = \begin{cases} 1 & \text{if the signal is transmitted} \\ 0 & \text{if the signal is not transmitted} \end{cases}$$

Thus the X_i are identically distributed, independent random variables, and by the strong law of large numbers,

$$\Pr \left\{ \lim_{k \rightarrow \infty} \frac{1}{k} \sum_{i=1}^k X_i = E\{X_i\} \right\} = 1 \quad (6.7)$$

But $E\{X_i\} = p(H_1)S$; therefore, the sequence $\{\hat{S}_k\}$ is an example of the required sequence $\{\phi_k\}$.

C. CONVERGENCE OF THE QUANTIZED SYSTEM

In the previous chapter we pointed out that in many cases the set Φ of all possible values of the parameter θ may not consist of a finite number of discrete points and the optimal learning system may not be realizable. In these cases under very general conditions (see Theorem 1,

Chapter V) a subset of Φ , say Φ_Q , exists and has the following properties:[†]

- (i) For every $\hat{\theta} \in \Phi$ and $\epsilon > 0$ there exists a $\theta_q \in \Phi_Q$ such that $\rho(d^*(\theta_q)|\hat{\theta}) \leq \rho(d^*(\hat{\theta})|\hat{\theta}) + \epsilon$.
- (ii) There are only a finite number of points $Q(\epsilon)$ in Φ_Q for any $\epsilon > 0$.

In this section we shall determine a sufficient condition that a system based on Φ_Q will converge in the following sense:

$$\lim_{k \rightarrow \infty} \rho_Q(\lambda_k|\hat{\theta}) = \rho(d^*(\theta_q)|\hat{\theta}) \quad (6.8)$$

with probability one, where $\rho_Q(\lambda_k|\hat{\theta})$ = average risk of system based on Φ_Q after k observations when $\hat{\theta}$ is the true value of θ , and

$$\rho(d^*(\theta_q)|\hat{\theta}) = \min_{\theta_1 \in \Phi_Q} \rho(d^*(\theta_1)|\hat{\theta}) \quad (6.9)$$

We shall show that this condition is met for most binary learning problems. Thus we will demonstrate that the suboptimum system is realizable and has a performance which is arbitrarily close to the performance of the optimum (unrealizable) system.

In order to determine a sufficient condition for convergence, we first note that the system based on Φ_Q computes the functions

$$\ell(X|\theta_j) \quad j = 1, 2, \dots, Q$$

[†]As previously defined, $\hat{\theta}$ is the true value of θ , and $\rho(d^*(\theta_q)|\hat{\theta})$ is the average risk of Bayes decision rule based on the assumption that θ_q is true when $\hat{\theta}$ is actually true.

and

$$\hat{P}(\theta_j | \lambda_k) = \frac{\prod_{i=1}^k p(x_i | \theta_j) P_o(\theta_j)}{\sum_{j=1}^Q \prod_{i=1}^k p(x_i | \theta_j) P_o(\theta_j)} \quad (6.10)$$

The system takes the sum of the products of $\ell(x | \theta_j)$ and $\hat{P}(\theta_j | \lambda_k)$ as the likelihood ratio:

$$\ell_Q(x | \lambda_k) = \sum_{j=1}^Q \hat{P}(\theta_j | \lambda_k) \ell(x | \theta_j) \quad (6.11)$$

The Bayes decision rule based on θ_q requires a comparison of $\ell(x | \theta_q)$ to a threshold. Thus if $\hat{P}(\theta_q | \lambda_k)$ converges to 1 when $\hat{\theta}$ is true, $\ell_Q(x | \lambda_k)$ will converge to $\ell(x | \theta_q)$ and the performance of the suboptimum system will converge to $\rho(d^*(\theta_q) | \hat{\theta})$. Theorem 2 states that if a minimum-risk solution exists, the system will converge to this solution.

Theorem 2: If there is a $\theta_q \in \Phi_Q$ such that

$$\rho(d^*(\theta_q) | \hat{\theta}) < \min_{\theta_j \in \Phi_Q} \rho(d^*(\theta_j) | \hat{\theta}) \quad (6.12)$$

and if the distribution of the observation under one hypothesis is independent of the unknown parameter θ [i.e., $p(x | \theta, H_2) = p(x | H_2)$], then

$$\lim_{k \rightarrow \infty} \hat{P}(\theta_q | \lambda_k) = 1 \quad \text{with probability one} \quad (6.13)$$

$$\lim_{k \rightarrow \infty} \hat{P}(\theta_j | \lambda_k) = 0 \quad \text{with probability one for all } \theta_j \in \Phi_Q, \theta_j \neq \theta_q$$

This theorem is proven in Appendix C. From the proof it is clear that when a unique minimum does not exist, then

$$\lim_{k \rightarrow \infty} \sum_{\theta_q \in \Phi_M} \hat{P}(\theta_q | \lambda_k) = 1 \quad \text{with probability one}$$

where Φ_M is the set of all points $\theta_q \in \Phi_Q$ which have minimum average risk. Since the points are equivalent from a performance standpoint, it makes no difference in performance whether the system converges so that

$$\ell_Q(X | \lambda_k) \rightarrow \ell(X | \theta_{q_1}) \quad \text{for some } \theta_{q_1} \in \Phi_M$$

or so that

$$\ell_Q(X | \lambda_k) \rightarrow \sum_{\theta_{q_1} \in \Phi_M} \ell(X | \theta_{q_1}) \hat{P}(\theta_{q_1} | \lambda_k)$$

Thus we may summarize by stating that if θ is an important parameter in the sense that knowing θ allows the design of a better system, then a system based on a discrete model for Φ will be finite, will exist, and will converge in performance.

D. RELATIONSHIP BETWEEN LEARNING WITH A TEACHER AND LEARNING WITHOUT A TEACHER

A very interesting relationship may be noted by referring back to Eq. (2.12). By writing the recursive form as a product, we find that

$$p(\theta | \lambda_{k-1}) = \prod_{i=1}^{k-1} \frac{\ell(X_i | \theta) + \alpha}{\ell(X_i | \lambda_{i-1}) + \alpha} p_o(\theta) \quad (6.14)$$

Thus Eq. (2.4) may be rewritten as

$$\ell(x_k | \lambda_{k-1}) = \int \ell(x_k | \theta) p_o(\theta) \prod_{i=1}^{k-1} \left[\frac{\ell(x_i | \theta) + \alpha}{\ell(x_i | \lambda_{i-1}) + \alpha} \right] d\theta \quad (6.15)$$

When a teacher is available, we may choose to train the system on the subset of the X which are known to contain a signal. In this case $p(H_1) = 1$, $p(H_2) = 0$; hence $\alpha = 0$, and the system computes a simpler form

$$\ell(x_k | \lambda_{k-1}) = \int \ell(x_k | \theta) p_o(\theta) \prod_{i=1}^{k-1} \left[\frac{\ell(x_i | \theta)}{\ell(x_i | \lambda_{i-1})} \right] d\theta \quad (6.16)$$

On the other hand, when we let $\alpha \rightarrow \infty$ the system becomes the usual nonlearning system

$$\ell(x_k | \lambda_{k-1}) = \ell(x_k) = \int \ell(x_k | \theta) p_o(\theta) d\theta \quad (6.17)$$

This is as it should be, since as $p(H_2) \rightarrow 1$, $p(H_1) \rightarrow 0$ and we cannot learn anything from the past.

Thus Eq. (2.12) describes a system applicable to all (parametrically expressible) binary decision problems. It applies even to those in which a learning sequence does not exist and to those in which a properly classified sequence does exist. For this reason the systems of Figs. 2 and 3 may be thought of as canonical decision systems. These figures provide the engineer with an insight into the relationship between the solutions to many binary decision problems, just as the tapped-delay-line canonical form of the linear filter provides an insight into linear filters.

E. SUMMARY OF CHAPTER VI

In this chapter we have applied the results of Appendix C and Refs. 15, 18, and 20 to demonstrate that the learning systems are stochastically stable and converge, and we have pointed out that the proposed systems are generally applicable to the entire parametric class of decision problems including the "no learning," "learning with a teacher," and "learning without a teacher" categories of problems.

We have also shown that in the cases where the unknown parameter is useful in the sense that knowledge of the parameter makes it possible to make more accurate decisions, a finite system always exists and converges in performance to a point arbitrarily close to the performance of a system with knowledge of the parameter.

Thus a system to learn without a teacher which has, from an engineering viewpoint, all of the properties of the optimum system may be constructed from a finite number of elements.

VII. SUMMARY OF RESULTS AND SUGGESTIONS FOR FUTURE WORK

A. RESULTS

The primary results of this investigation have been summarized in detail at the end of pertinent chapters, and are briefly described in the form of the following four major contributions of this work. In the first two items, a statistical model has been obtained which fits a large class of interesting decision problems, and a method has been developed to solve these problems. The third and fourth contributions have been related to the practicality of the theoretical systems.

1. A recursive relation has been developed which describes the structure of learning systems which are optimum for any length of learning sequence. The problems which may be solved by such systems are restricted to the parametric class of decision problems in which the functional form of the underlying probability measures is known; however this class includes problems in which the learning sequence is not previously classified, as well as problems in which the a priori probability of occurrence of different classes of observations is unknown.
2. The solution has been extended to problems in which the unknown parameter is a time-varying random variable. It has been shown that solutions to the time-varying problem are straightforward modifications of solutions to fixed parameter problems.

Thus we have obtained a statistical model which fits a large class of interesting decision problems and have developed a method to solve these problems. The method results in a theoretical and functional description of decision systems to solve the problems. Our third and fourth contributions have been related to the practicality of the theoretical systems.

3. It has been demonstrated that in the case where the unknown parameter may take on only a finite number of values, the optimum learning system requires a finite memory and is therefore realizable with a finite number of elements.
4. It has also been demonstrated that so long as the underlying probability measures are either discrete or absolutely continuous in the observation space, and so long as the Bayes decision rule depends upon the unknown parameter, a finite-memory suboptimum system exists which has performance arbitrarily close to the performance of the optimum system.

B. PROBLEMS FOR ADDITIONAL RESEARCH

There are many interesting and important applications of decision machines which learn, just as there are many important general problems involving such machines. Some of the more outstanding problems are given below as suggested areas for future research:

1. It is clear that in many applications the functional form of the underlying probability measures is unknown, and thus many problems may not be treated as parametric learning problems. A systematic technique for the solution of such problems would be extremely useful, and an investigation of the possibility of treating such problems by expanding the probability measures in a series of known functions with unknown parameters and coefficients might lead to such a technique.
2. In this study a finite-memory system which is optimum in an engineering sense has been found by approximating the space of the unknown parameter with a discrete space. An investigation of the structure of the optimum system under a finite-memory constraint might lead to additional insight into the solution of learning problems.
3. The investigation of performance bounds has been incomplete and the bounds determined have been undesirably loose. This is due primarily to the fact that such bounds depend very much on the particular learning problem being solved. It is presently necessary to apply difficult, time-consuming numerical computation techniques or to build or simulate the system in order to determine whether the resulting performance will be acceptable or to compare the optimum system with some suboptimum system. It seems clear that a simpler procedure for obtaining tighter bounds on performance would be very useful.

APPENDIX A. EVALUATION OF $P(B_k)$

In order to evaluate $P(B_k)$ we first note that B_k may occur only when $P(\hat{S}|\lambda_k) \leq 1/2$; therefore $P(B_k) \leq \Pr \{P(\hat{S}|\lambda_k) \leq 1/2\}$. To evaluate this bound we shall determine bounds on the moments of the distribution of the random variable $P(\hat{S}|\lambda_k)$ and apply a Tchebysheff type of bound. Thus in any particular case the resulting bound may be very loose; however for the example of Chapter III it is clear that the bound is a useful one.

Consider the estimate of $\hat{S}$ given by

$$S_k^* = \sum_{i=1}^m s_i P(s_i|\lambda_k) \quad (A.1)$$

Then

$$E\left\{|\hat{S} - S_k^*|^2\right\} \leq E\left\{|\hat{S} - \phi_k|^2\right\} \quad (A.2)$$

for any other estimate ϕ_k based on λ_k because S_k^* is the least-mean-square error estimate of $\hat{S}$ based on λ_k . In particular, consider the estimate

$$\phi_k = \frac{1}{kp_1} \sum_{i=1}^k X_i \quad (A.3)$$

Now, if

$$(i) \quad X = \begin{cases} \hat{S} + N & \text{with probability } p_1 \\ N & \text{with probability } p_2 = 1-p_1 \end{cases}$$

(ii) Signal and noise are independent,

(iii) $E\{\hat{S}\} = E\{N\} = 0$, and

(iv) The noise is bandlimited and white with variance σ_n^2 ,

then

$$E\{(\hat{S} - \phi_k)_t (\hat{S} - \phi_k)\} = \frac{\sigma_n^2}{kp_1^2} + \frac{p_1 p_2}{kp_1^2} E\{\hat{S}_t \hat{S}\} \quad (A.4)$$

and therefore

$$E\{(\hat{S} - S_k^*)_t (\hat{S} - S_k^*)\} \leq \frac{\sigma_n^2 + p_1 p_2 E\{\hat{S}_t \hat{S}\}}{kp_1^2} \quad (A.5)$$

We may evaluate $E\{(\hat{S} - S_k^*)_t (\hat{S} - S_k^*)\}$ as follows. From (A.1) and the fact that $(S_{i_t} S_j)/\sigma_n^2 = \delta_{ij} R$ [see text following Eq. (3.6)], we have that

$$E\{\hat{S}_t S_k^*\} = E\{\hat{S}_t \hat{S}\} p(S|\lambda_k) \quad (A.6)$$

and

$$E\{S_{k_t}^* S_k^*\} = E\left\{\left[\sum_{i=1}^m S_i P(S_i|\lambda_k)\right]_t \left[\sum_{i=1}^m S_i P(S_i|\lambda_k)\right]\right\} \quad (A.7)$$

Because of the symmetry, $E\{[P(S_i|\lambda_k)]^2\}$ is constant for all $S_i \neq \hat{S}$, so that

$$E\{S_{k_t}^* S_k^*\} = E\{\hat{S}_t \hat{S}\} \left[E\{[P(\hat{S}|\lambda_k)]^2\} + (m-1) E\{[P(S_i|\lambda_k)]^2\} \right] \quad (A.8)$$

By factoring and collecting terms we have

$$E\{(\hat{S} - S_k^*)_t (\hat{S} - S_k^*)\} = E\{\hat{S}_t \hat{S}\} \left[E\{(\hat{P}_k - 1)^2\} + (m-1) E\{P_{ik}^2\} \right] \quad (A.9)$$

where

$$\hat{p}_k = P(\hat{S}|\lambda_k) \quad (\text{A.10a})$$

$$p_{ik} = P(S_i|\lambda_k) \quad (\text{A.10b})$$

Now $(m-1) E\{p_{ik}^2\} \geq 0$ and is small, so that

$$E\{(1 - \hat{p}_k)^2\} \leq \frac{\sigma_n^2 + p_1 p_2 E\{\hat{S}_t \hat{S}\}}{kp_1^2 E\{\hat{S}_t \hat{S}\}} \quad (\text{A.11})$$

or

$$E\{(1 - \hat{p}_k)^2\} \leq \frac{\frac{1}{R} + p_1 p_2}{kp_1^2} \quad (\text{A.12})$$

where $R = E\{\hat{S}_t \hat{S}\} / \sigma_n^2$.

In order to obtain a bound on $P(B_k)$ we also require the first moment of $\hat{p}_k$. To bound this we write

$$\sum_{i=1}^m p_{ik} = 1 \quad \text{therefore} \quad \sum_{\substack{i=1 \\ i \neq \hat{i}}}^m p_{ik} = 1 - \hat{p}_k \quad (\text{A.13})$$

Hence

$$\left(\sum_{\substack{i=1 \\ i \neq \hat{i}}}^m p_{ik} \right)^2 = (1 - \hat{p}_k)^2 \quad (\text{A.14})$$

But

$$\left(\sum_{\substack{1 \\ i \neq 1}}^m p_{ik} \right)^2 = \sum_{\substack{1 \\ i \neq 1}}^m p_{ik}^2 + \sum_{i \neq 1} \sum_{j \neq 1} p_{ik} p_{jk} \quad (\text{A.15})$$

so that

$$E\{(1 - \hat{p}_k)^2\} = (m-1) E\{p_{ik}^2\} + (m-1)(m-2) E^2\{p_{ik}\} \quad (\text{A.16})$$

From Eqs. (A.5) and (A.6) we have

$$E\{(1 - \hat{p}_k)^2\} \leq \frac{1}{kp_1^2 R'} - (m-1) E\{p_{ik}^2\} \quad (\text{A.17})$$

where

$$R' = \frac{R}{1 + p_1 p_2 R} \quad (\text{A.18})$$

so that

$$(m-1)(m-2) E^2\{p_{ik}\} \leq \frac{1}{kp_1^2 R'} - 2(m-1) E\{p_{ik}^2\} \quad (\text{A.19})$$

But since

$$\text{var}(p_{ik}) = E\{p_{ik}^2\} - E^2\{p_{ik}\} \geq 0 \quad (\text{A.20a})$$

$$E\{p_{ik}^2\} \geq E^2\{p_{ik}\} \quad (\text{A.20b})$$

we can write

$$(m-1)^2 E^2\{P_{1k}\} \leq \frac{1}{kp_1^2 R'} \quad (A.21)$$

Finally, by utilizing (A.13) we may bound the first moment as

$$E\{\hat{P}_k\} \leq 1 - \left(\frac{m-1}{mkp_1^2 R'} \right)^{1/2} \quad (A.22)$$

By utilizing this first-moment bound and the previous bound on the second moment and applying a Tchebysheff-type bound [Ref. 34, p. 93], we have

$$\Pr \left\{ \hat{P}_k \leq \frac{1}{2} \right\} \leq \frac{4}{\left\{ (p_1 k R' + 8) - 4 \left[\frac{(m-1)k R'}{m} \right]^{1/2} \right\} p_1} \quad (A.23)$$

which is valid so long as $kp_1^2 R' > 16$. By using (A.18) we have

$$\Pr \left\{ \hat{P}_k \leq \frac{1}{2} \right\} \leq \frac{4(1 + p_1 p_2 R)}{p_1 \left\{ p_1 k R + 8(1 + p_1 p_2 R) - 4 \left[\frac{(m-1)k R (1 + p_1 p_2 R)}{m} \right]^{1/2} \right\}} \quad (A.24)$$

Because $p_1 k R \gg 8(1 + p_1 p_2 R)$ for large k , we may take as our bound for $p_1 P(B_k)$,

$$p_1 P(B_k) \leq \frac{4(1 + p_1 p_2 R)}{p_1 k R - 4 \left[\frac{(m-1)k R (1 + p_1 p_2 R)}{m} \right]^{1/2}} \quad (A.25)$$

APPENDIX B. PROOF OF THEOREM 1

For convenience in presenting this proof, Theorem 1 of Chapter V is repeated below.

Theorem 1. Designate by Φ the space of all possible values of the vector parameter θ , and let the range of each coordinate of θ be bounded. Then if $p(X|\theta, H_2)$ is independent of θ , and $p(X|\theta, H_1)$ is a continuous function of θ for all $\theta \in \Phi$ and all X , there exists a subset of Φ , say $\Phi_Q = \{\theta_1, \theta_2, \dots, \theta_Q\}$, with a finite number, Q , of discrete values of θ such that for any $\epsilon > 0$ and all $\hat{\theta} \in \Phi$ there is a $\theta_q \in \Phi_Q$ which satisfies

$$\rho(d^*(\theta_q)|\hat{\theta}) \leq \rho(d^*(\hat{\theta})|\hat{\theta}) + \epsilon$$

where $\rho(d^*(\theta_q)|\hat{\theta})$ is the average risk of the Bayes decision rule based on the assumption that θ_q is true ($d^*(\theta_q)$) when $\hat{\theta}$ is true.

Proof. If $p(X|\theta, H_1)$ is a continuous function of θ , then so also is the integral over any range of X . That is

$$\int_R p(X|\theta, H_1) dX$$

is a continuous function of θ . Therefore, given any $\epsilon' > 0$, there is a $\delta > 0$ such that if θ_i and θ_j both lie within a sphere of radius δ , then

$$\left| \int_R p(X|\theta_i, H_1) dX - \int_R p(X|\theta_j, H_1) dX \right| < \epsilon'$$

We may therefore choose as a possible set $\{\Phi_Q\}$, all the points in Φ which are distance δ along some coordinate from an arbitrary point. Since the range of values of each coordinate is bounded, there will be

only a finite number of points in this set; furthermore for every $\hat{\theta}$ in Φ there will be a member, say θ_q , in Φ_Q such that

$$\left| \int_R p(X|\theta_q, H_1) dX - \int_R p(X|\hat{\theta}, H_1) dX \right| < \epsilon'$$

The Bayes decision rule based on θ_q , $d^*(\theta_q)$, divides the observation space into two mutually exclusive regions R_q and $\bar{R}_q$. If $X \in R_q$, then decision rule $d^*(\theta_q)$ results in the decision to accept hypothesis 1. If $X \in \bar{R}_q$, then the decision is to accept H_2 . Hence the average risk $\rho(d^*(\theta_q)|\theta_q)$ when θ_q is true is given by

$$\rho(d^*(\theta_q)|\theta_q) = p_1 P(X \in \bar{R}_q | \theta_q, H_1) + L p_2 P(X \in R_q | \theta_q, H_2)$$

But $P(X \in R_q | \theta_q, H_2) = P(X \in R_q | H_2)$ because the distribution of X is independent of θ when H_2 is true. Hence we must have

$$p_1 P(X \in \bar{R}_q | \hat{\theta}, H_1) + L p_2 P(X \in R_q | H_2) - p_1 \epsilon' < \rho(d^*(\theta_q)|\theta_q)$$

and

$$\rho(d^*(\theta_q)|\theta_q) < p_1 P(X \in \bar{R}_q | \hat{\theta}, H_1) + L p_2 P(X \in R_q | H_2) + p_1 \epsilon'$$

or

$$\left| \rho(d^*(\theta_q)|\theta_q) - \rho(d^*(\theta_q)|\hat{\theta}) \right| < p_1 \epsilon' \quad (B.1)$$

Similarly, by starting with $\rho(d^*(\hat{\theta})|\hat{\theta})$ we have

$$\left| \rho(d^*(\hat{\theta})|\hat{\theta}) - \rho(d^*(\hat{\theta})|\theta_q) \right| < p_1 \epsilon' \quad (B.2)$$

Because d^* is a Bayes rule, the following must hold:

$$\rho(d^*(\theta_q) | \theta_q) \leq \rho(d^*(\hat{\theta}) | \theta_q) \quad (B.3)$$

$$\rho(d^*(\hat{\theta}) | \hat{\theta}) \leq \rho(d^*(\theta_q) | \hat{\theta}) \quad (B.4)$$

By inserting (B.2) in (B.3) we obtain

$$\rho(d^*(\theta_q) | \theta_q) \leq \rho(d^*(\hat{\theta}) | \hat{\theta}) + p_1 \epsilon' \quad (B.5)$$

and inserting (B.1) in (B.4) we obtain

$$\rho(d^*(\hat{\theta}) | \hat{\theta}) \leq \rho(d^*(\theta_q) | \theta_q) + p_1 \epsilon' \quad (B.6)$$

so that

$$\left| \rho(d^*(\theta_q) | \theta_q) - \rho(d^*(\hat{\theta}) | \hat{\theta}) \right| \leq p_1 \epsilon' \quad (B.7)$$

The combination of (B.1) and (B.7) yields

$$\left| \rho(d^*(\theta_q) | \hat{\theta}) - \rho(d^*(\hat{\theta}) | \hat{\theta}) \right| \leq 2p_1 \epsilon'$$

and thus by choosing $\epsilon' < \epsilon / 2p_1$ we have proven the theorem.

APPENDIX C. PROOFS OF STABILITY AND CONVERGENCE

1. System Stability

In order to prove that the system is stable we first prove a more general theorem[†] regarding a property of the probability measure $P_k(\theta) = P(\theta|X_1, \dots, X_k)$, which is the cumulative distribution function corresponding to the density $p(\theta|\lambda_k)$.

Theorem 3. Any sequence $\{g_1, g_2, \dots, g_{n+1}\}$ such that

$$g_k = \int_{\Phi} f(\theta) dP_k(\theta) \quad (C.1)^{\ddagger}$$

where

$$P_k(\theta) = P(\theta|X_1, \dots, X_k) \quad 1 \leq k \leq n+1 \quad (C.2)$$

is a bounded martingale if

- (i) $f(\theta)$ is any nonnegative Lebesgue measurable function,
- (ii) $\max f(\theta) = M < \infty$.

Proof. A martingale is defined [Ref. 35, p. 293] as a sequence of random variables $\{X_1, X_2, \dots, X_n, z\}$ such that

- (iii) $E\{|z|\} < \infty$,
- (iv) $X_n = E\{z|w_1, w_2, \dots, w_n\}$ for some set of random variables $\{w_i\}$.

Thus to prove the martingale property, it is sufficient to prove

- a. $E\{|g_{n+1}|\} < \infty$
- b. $g_n = E\{g_{n+1}|X_1, \dots, X_n\}$

[†]This theorem is due to Daly [Ref. 20]; the proof is repeated for convenience.

[‡]In order to include the case where $P_k(\theta)$ is a step function, the integral here is meant in the Lebesgue-Stieltjes sense (see, e.g., Ref. 1).

First we prove (a). Since $f(\theta)$ is nonnegative and bounded by M on Φ , and $\int dP_{n+1}(\theta) = 1$, then g_{n+1} is nonnegative and bounded by M ; i.e.,

$$0 \leq g_{n+1} = \int f(\theta) dP_{n+1}(\theta) \leq \int M dP_{n+1}(\theta) = M \int dP_{n+1}(\theta) = M < \infty \quad (C.3)$$

hence

$$|g_{n+1}| \leq M < \infty \quad (C.4)$$

and

$$E\{|g_{n+1}|\} \leq M < \infty \quad (C.5)$$

Since this is true for all n , we also have

$$\lim_{n \rightarrow \infty} E\{|g_n|\} \leq \lim_{n \rightarrow \infty} M = M < \infty \quad (C.6)$$

This relation will be required in the proof of the boundedness of the sequence.

To prove (b) we must show[†]

$$E\left\{\int_{\Phi} f(\theta) p(\theta|X_1, \dots, X_{n+1}) d\theta \mid X_1, \dots, X_n\right\} = \int_{\Phi} f(\theta) p(\theta|X_1, \dots, X_n) d\theta \quad (C.7)$$

where the expectation is over the space χ of X_{n+1} . We may write

$$E\{g_{n+1} | X_1, \dots, X_n\} = \int_{\chi} \left[\int_{\Phi} f(\theta) p(\theta|X_1, \dots, X_{n+1}) d\theta \right] \cdot p(X_{n+1} | X_1, \dots, X_n) dX_{n+1} \quad (C.8)$$

[†]In this case, since we are only interested in finite n , we need not contend with step functions, hence we write $p_k(\theta) = dP_k(\theta)/d\theta$ for easier manipulation.

Interchanging the order of integration over Φ and χ , we have:

$$\begin{aligned}
 E\{g_{n+1}|X_1, \dots, X_n\} &= \int_{\Phi} f(\theta) \left[\int_{\chi} \frac{p(\theta|X_1, \dots, X_n) p(X_1, \dots, X_n|X_{n+1})}{p(X_1, \dots, X_n)} \right. \\
 &\quad \left. \cdot P(X_{n+1}) dX_{n+1} \right] d\theta \\
 &= \int_{\Phi} f(\theta) p(\theta|X_1, \dots, X_n) d\theta \\
 &= g_n
 \end{aligned} \tag{C.9}$$

Thus the sequence $\{g_n; n = 1, 2, \dots\}$ is a martingale. Doob [Ref. 35, p. 319] shows in theorem VII, 4.1 that if the sequence $\{g_n; n \geq 1\}$ is a martingale, and if $\lim E\{|g_n|\} = M < \infty$, then $\lim_{n \rightarrow \infty} g_n = g_{\infty}$ exists with probability one. Thus the sequence $\{g_n; n \geq 1\}$ does indeed converge to a limit with probability one.

This theorem is directly applicable to the proof of system stability. We make the identification $f(\theta) = \ell(X|\theta)$. Then $f(\theta)$ will be a non-negative Lebesgue measurable function of θ . If in addition $\ell(X|\theta)$ is a bounded function of θ for all X , then the sequence $\ell(X_k|\lambda_{k-1})$ is a bounded martingale and $\lim_{k \rightarrow \infty} \ell(X_k|\lambda_{k-1}) < \infty$ with probability one.

2. Convergence of the Optimal System

In order to find the limit to which the system converges we first state a theorem due to Braverman [Ref. 15].

Theorem 4. If there exists a sequence of functions $\{\phi_k(X_1, \dots, X_k)\}$ such that $\lim_{k \rightarrow \infty} \phi_k = \hat{\theta}$ with probability one, where $\hat{\theta}$ is the true value of θ , then

$$\lim_{k \rightarrow \infty} P(\theta|X_1, \dots, X_k) = \begin{cases} 1 & \theta \cong \hat{\theta} \\ 0 & \theta < \hat{\theta} \end{cases} \tag{C.10}$$

By $\theta < \hat{\theta}$ we mean that every coordinate of θ is less than every coordinate of $\hat{\theta}$ since θ may be a vector-valued parameter.

Proof. Braverman proves this theorem for the case of learning with a teacher, drawing on the fact that if the sequence $\{X_1'\}$ is known to arise from a particular class, then

$$g_k' = \int_{\Phi} g(\theta) dP_k'(\theta) \quad (C.11)$$

is a bounded martingale if $g(\theta)$ is bounded and Lebesgue measurable on Φ .

We have already proven that g_k is a bounded martingale even when $\{X_1\}$ does not arise from a single class. Thus if we consider the sequence of functions

$$\left\{ P_k(E_{\theta}) = \int_{E_{\theta}} dP_k(\theta) \right\} \quad (C.12)$$

this sequence will be a bounded martingale because it can be written as

$$P_k(E_{\theta}) = \int_{\Phi} I_{E_{\theta}} dP_k(\theta) \quad (C.13)$$

where

$$I_{E_{\theta}} = \begin{cases} 1 & \theta \in E_{\theta} \\ 0 & \theta \notin E_{\theta} \end{cases} \quad (C.14)$$

is the indicator function of the set $\{E_{\theta}\}$; hence

$$\lim_{k \rightarrow \infty} P_k(E_{\theta}) = P_{\infty}(E_{\theta}) \quad \text{with probability one} \quad (C.15)$$

Loève [Ref. 36] points out that if the sequence $\{X_1, X_2, \dots, X_n, z\}$ is a bounded martingale, then $E\{z|X_1, \dots, X_n\}$ converges with probability one to z . If we let $z = I_{E_\theta}$, then the sequence $\{P_k(E_\theta)\}$ must converge to either 1 or 0.

The existence of the convergent sequence $\{\phi_k(\lambda_k)\}$ must imply that $P_k(E_\theta)$ converges to 1 when $\hat{\theta}$ is contained in E_θ , and converges to 0 when $\hat{\theta}$ is not in E_θ .

Thus $P_\infty(\theta)$ must be a (multidimensional) step function with a discontinuity at $\hat{\theta}$ with probability one.

We may extend this theorem to the following corollary.

Corollary. If there exists a sequence $\{\phi_k(X_1, \dots, X_k)\}$ such that

$$\lim_{k \rightarrow \infty} \phi_k = \hat{\theta} \quad \text{with probability one} \quad (\text{C.16})$$

then

$$\lim_{k \rightarrow \infty} \int_{\Phi} f(\theta) dP_k(\theta) = f(\hat{\theta}) \quad \text{with probability one} \quad (\text{C.17})$$

if $f(\theta)$ is continuous on Φ .

This follows from the above theorem and the fact that

$$\lim_{k \rightarrow \infty} \int_{\Phi} f(\theta) dP_k(\theta) = \int_{\Phi} f(\theta) dP_\infty(\theta) \quad \text{with probability one} \quad (\text{C.18})$$

if $f(\theta)$ is continuous on Φ and $P_\infty(\theta)$ has bounded variation on Φ . By definition of the Lebesgue-Stieltjes integral, if $P_\infty(\theta)$ is a step function at $\hat{\theta}$, then

$$\int_{\Phi} f(\theta) dP_\infty(\theta) = f(\hat{\theta}) \quad (\text{C.19})$$

Hence if $\ell(X|\theta)$ is a continuous function of θ , we have the fact that

$$\lim_{k \rightarrow \infty} \int_{\Phi} \ell(X|\theta) dP_k(\theta) = \ell(X|\hat{\theta}) \quad \text{with probability one} \quad (\text{C.20})$$

(where $\hat{\theta}$ is the true value of θ) if the sequence $\{\phi_k\}$ exists.

3. Convergence of the Quantized System

In order to prove Theorem 2, Chapter V we first determine a sufficient condition for convergence as follows.

Theorem 5. If

$$E\{\log p(X|\theta_q) - \log p(X|\theta_j) | \hat{\theta}\} > 0 \quad (\text{C.21})$$

for some $\theta_q \in \Phi_Q$ and every $\theta_j \in \Phi_Q$, $\theta_j \neq \theta_q$, then

$$\lim_{k \rightarrow \infty} \hat{P}(\theta_q | \lambda_k) = 1 \quad \text{with probability one} \quad (\text{C.22a})$$

$$\lim_{k \rightarrow \infty} \hat{P}(\theta_j | \lambda_k) = 0 \quad \text{with probability one} \quad (\text{C.22b})$$

Proof. If

$$E\{\log p(X|\theta_q) - \log p(X|\theta_j)\} = \beta > 0 \quad (\text{C.23})$$

for all $\theta_j \in \Phi_Q$, $\theta_j \neq \theta_q \in \Phi_Q$, then

$$\lim_{k \rightarrow \infty} \sum_{i=1}^k \log \left[\frac{p(X_i | \theta_q)}{p(X_i | \theta_j)} \right] = k\beta \quad \text{with probability one} \quad (\text{C.24})$$

That is, for every $\epsilon > 0$, there exists a k such that

$$\Pr \left\{ \text{LUB} \left| \sum_{i=1}^k \log \left[\frac{p(X_i | \theta_q)}{p(X_i | \theta_j)} \right] - k\beta \right| < \epsilon \right\} = 1 \quad (\text{C.25})$$

where LUB means least upper bound. But

$$\text{LUB} \left| \sum_{i=1}^k \log \left[\frac{p(X_i | \theta_q)}{p(X_i | \theta_j)} \right] - k\beta \right| < \epsilon \quad (\text{C.26})$$

implies that

$$\text{LUB} \left| \exp \left\{ - \sum_{i=1}^k \log \left[\frac{p(X_i | \theta_q)}{p(X_i | \theta_j)} \right] \right\} - \exp(-k\beta) \right| < e^{-k\beta}(e^\epsilon - 1) \quad (\text{C.27})$$

Therefore, for all $\delta > 0$, there exists a k such that

$$\Pr \left\{ \text{LUB} \left| \prod_{i=1}^k \frac{p(X_i | \theta_j)}{p(X_i | \theta_q)} - e^{-k\beta} \right| < \delta \right\} = 1 \quad (\text{C.28})$$

or, for every $\delta' > 0$ there exists a k such that

$$\Pr \left\{ \text{LUB} \left| \prod_{i=1}^k \frac{p(X_i | \theta_j)}{p(X_i | \theta_q)} \right| < \delta' \right\} = 1 \quad (\text{C.29})$$

The function being computed is $\hat{P}(\theta_j | \lambda_k)$ which may be written as below.

$$\hat{P}(\theta_j | \lambda_k) = \frac{\prod_{i=1}^k \frac{p(X_i | \theta_j)}{p(X_i | \theta_q)} \frac{P_o(\theta_j)}{P_o(\theta_q)}}{1 + \sum_{j=1}^Q \prod_{i=1}^k \frac{p(X_i | \theta_j)}{p(X_i | \theta_q)} \frac{P_o(\theta_j)}{P_o(\theta_q)}} \quad (C.30)$$

For each $\theta_j \in \Phi_Q$ and each k , define ζ_{jk} such that

$$\text{LUB} \prod_{i=1}^k \frac{p(X_i | \theta_j)}{p(X_i | \theta_q)} < \zeta_{jk} \quad (C.31)$$

then

$$\text{LUB} |\hat{P}(\theta_j | \lambda_k)| < \frac{\zeta_{jk} \frac{P_o(\theta_j)}{P_o(\theta_q)}}{1 + \sum_{j=1}^Q \prod_{i=1}^k \frac{p(X_i | \theta_j)}{p(X_i | \theta_q)} \frac{P_o(\theta_j)}{P_o(\theta_q)}} < \zeta_{jk} \frac{P_o(\theta_j)}{P_o(\theta_q)} \quad (C.32)$$

Hence for every $\zeta > 0$ there exists a k such that

$$\text{Pr} \{ \text{LUB} |\hat{P}(\theta_j | \lambda_k)| < \zeta \} = 1 \quad (C.33)$$

Similarly,

$$\text{LUB} |1 - \hat{P}(\theta_q | \lambda_k)| < 1 - \left[1 + \sum_{j=1}^Q \prod_{i=1}^k \frac{p(X_i | \theta_j)}{p(X_i | \theta_q)} \frac{P_o(\theta_j)}{P_o(\theta_q)} \right]^{-1} < \sum_{j=1}^Q \zeta_{jk} \frac{P_o(\theta_j)}{P_o(\theta_q)} \quad (C.34)$$

so that by choosing

$$\max_j \zeta_{jk} < \frac{\zeta}{Q-1} \frac{P_o(\theta_q)}{P_o(\theta_j)}$$

we obtain

$$\text{LUB } |1 - \hat{P}(\theta_q | \lambda_k)| < \zeta \quad (\text{C.35})$$

Therefore, for all $\zeta > 0$, there exists a k such that

$$\Pr \left\{ \text{LUB } |\hat{P}(\theta_q | \lambda_k) - 1| < \zeta \right\} = 1 \quad (\text{C.36})$$

So that

$$\lim_{k \rightarrow \infty} \hat{P}(\theta_q | \lambda_k) = 1 \quad \text{with probability one} \quad (\text{C.37a})$$

$$\lim_{k \rightarrow \infty} \hat{P}(\theta_j | \lambda_k) = 0 \quad \text{with probability one} \quad (\text{C.37b})$$

which proves Theorem 5.

Theorem 2. If there is a $\theta_q \in \Phi_Q$ such that

$$\rho(d^*(\theta_q) | \hat{\theta}) < \min_{\theta_j \in \Phi_Q} \rho(d^*(\theta_j) | \hat{\theta}) \quad (\text{6.12})$$

and if the distribution of the observation under one hypothesis is independent of the unknown parameter θ [i.e., $p(X|\theta, H_2) = p(X|H_2)$], then

$$\lim_{k \rightarrow \infty} \hat{P}(\theta_q | \lambda_k) = 1 \quad \text{with probability one} \quad (\text{6.13})$$

$$\lim_{k \rightarrow \infty} \hat{P}(\theta_j | \lambda_k) = 0 \quad \text{with probability one for all } \theta_j \in \Phi_Q, \theta_j \neq \theta_q$$

Proof. Condition (6.12) implies that

$$\int_a^\infty p_q(\ell) d\ell > \int_a^\infty p_j(\ell) d\ell \quad (C.38)$$

where $p_q(\ell)$ = probability density of $\ell(X|\theta_q)$

a = any real number > 0

To prove this, we observe that (6.12) implies

$$\begin{aligned} Lp_2 \int_\gamma^\infty p_q(\ell|H_2) d\ell + p_1 \int_{-\infty}^\gamma p_q(\ell|H_1) d\ell &< Lp_2 \int_\gamma^\infty p_j(\ell|H_2) d\ell \\ &+ p_1 \int_{-\infty}^\gamma p_j(\ell|H_1) d\ell \end{aligned} \quad (C.39)$$

where $\gamma = Lp_2/p_1$

p_i = a priori probability of H_i being true

By rearranging this inequality, and changing limits, we have

$$p_1 \int_\gamma^\infty [p_q(\ell|H_1) - p_j(\ell|H_1)] d\ell > Lp_2 \int_\gamma^\infty [p_q(\ell|H_2) - p_j(\ell|H_2)] d\ell \quad (C.40)$$

for all $\gamma > 0$. Now assume that

$$\int_{a_0}^\gamma p_q(\ell) d\ell \leq \int_{a_0}^\gamma p_j(\ell) d\ell \quad (C.41)$$

for some real number a_o . Then

$$\int_{a_o}^{\infty} [p_1 p_q(\ell|H_1) + p_2 p_q(\ell|H_2)] d\ell \leq \int_{a_o}^{\infty} [p_1 p_j(\ell|H_1) + p_2 p_j(\ell|H_2)] d\ell \quad (C.42)$$

or

$$p_1 \int_{a_o}^{\infty} [p_q(\ell|H_1) - p_j(\ell|H_1)] d\ell \leq - p_2 \int_{a_o}^{\infty} [p_q(\ell|H_2) - p_j(\ell|H_2)] d\ell \quad (C.43)$$

Combining (C.40) and (C.43) yields

$$L p_2 \int_{\gamma}^{\infty} [p_q(\ell|H_2) - p_j(\ell|H_2)] d\ell < - p_2 \int_{a_o}^{\infty} [p_q(\ell|H_2) - p_j(\ell|H_2)] d\ell \quad (C.44)$$

for all $\gamma > 0$. Suppose that $a_o > 0$, then we can choose $\gamma = a_o$, hence $L p_2 = a_o p_1$, and (C.44) becomes

$$a_o p_1 < - p_2$$

Hence a_o cannot be positive, and for all positive real numbers a , (C.38) must hold.

Consider the function

$$E\{\log p(X|\theta_q) - \log p(X|e_j) | \hat{e}\} = E_{q,j}$$

It may be written as

$$E_{q,j} = E\left\{\log [\alpha + \ell(X|\theta_q)] - \log [\alpha + \ell(X|\theta_j)] \middle| \hat{\theta}\right\} \\ + E\left\{\log [p_1 p(X|\theta_q, H_2)] - \log [p_1 p(X|\theta_j, H_2)] \middle| \hat{\theta}\right\}$$

where $\alpha = p_2/p_1$. But since $p(X|\theta_q, H_2) = p(X|\theta_j, H_2) = p(X|H_2)$, the second term on the right is zero. The function $\log [\alpha + \ell]$; $\ell \geq 0$ is monotonically increasing and continuous; hence it may be approximated by a sum of simple functions:

$$\lim_{\substack{N \rightarrow \infty \\ \Delta \rightarrow 0}} \sum_{i=0}^N \beta_i \phi_i(\ell) = \log [\alpha + \ell] \quad \text{almost everywhere}$$

where

$$\phi_i(\ell) = \begin{cases} 1 & \text{when } \ell \geq i\Delta \\ 0 & \text{when } \ell < i\Delta \end{cases}$$

$$\beta_i = \log [\alpha + i\Delta] - \log [\alpha + (i-1)\Delta]$$

so that

$$E_{q,j} = \int_0^\infty \log [\alpha + \ell] \{p_q(\ell) - p_j(\ell)\} d\ell \\ = \int_0^\infty \lim_{\substack{N \rightarrow \infty \\ \Delta \rightarrow 0}} \sum_{i=1}^N \beta_i \phi_i(\ell) \{p_q(\ell) - p_j(\ell)\} d\ell$$

$$= \lim_{\substack{N \rightarrow \infty \\ \Delta \rightarrow 0}} \sum_{i=1}^N \beta_i \int_{i\Delta}^{\infty} (p_q(\ell) - p_j(\ell)) d\ell$$

But the sum of a set of positive numbers must be positive so that

$$E(\log p(X|\theta_q) - \log p(X|\theta_j)) > 0$$

By Theorem 5 this implies the convergence of $\hat{P}(\theta_q|\lambda_k)$ to 1, and proves Theorem 2.

REFERENCES

1. J. C. Burkill, The Lebesgue Integral, Cambridge University Press, London, 1951.
2. D. Middleton, An Introduction to Statistical Communication Theory, McGraw-Hill Book Co., Inc., New York, 1960.
3. A. Wald, Statistical Decision Functions, John Wiley & Sons, New York, 1950.
4. N. Abramson and J. Farison, "On Applied Decision Theory," Rept. SEL-62-095 (TR No. 2005-2), Stanford Electronics Laboratories, Stanford, Calif., Sep 1962.
5. R. Price, "Optimum Detection of Random Signals in Noise, with Applications to Scatter-Multipath Communication," Part I, Trans. IRE, PGIT-2, Dec 1956, pp. 125-135.
6. R. Price and P. E. Green, Jr., "A Communication Technique for Multipath Channels," Proc. IRE, 46, 1958, pp. 555-570.
7. T. Kailath, "Adaptive Matched Filters," in Mathematical Optimization Techniques, edited by R. Bellman, University of California Press, Berkeley, 1963, pp. 109-140.
8. T. Kailath, "Optimum Receivers for Randomly Varying Channels," Proc. Fourth London Symposium on Information Theory, Butterworth Scientific Press, London, 1961, pp. 109-124.
9. J. G. Proakis and R. Drouilhet, Jr., "Performance of Coherent Detection Systems Using Decision Directed Channel Measurement," Rept. 64 G-1, Lincoln Laboratory, MIT, Cambridge, Mass., 27 Jun 1963.
10. H. J. Scudder III, "Adaptive Communication Receivers," Rept. No. 64-3, Electronic Research Laboratory, University of California, Berkeley, Apr 1964.
11. E. M. Glaser, "Signal Detection by Adaptive Filters," Trans. IRE, PGIT-7, Apr 1961, pp. 87-98.
12. C. V. Jackowatz, R. L. Shuey, and G. M. White, "Adaptive Waveform Recognition," Proc. Fourth London Symposium on Information Theory, Butterworth Scientific Press, London, 1961, pp. 317-326.
13. M. J. Hinich, "A Model for a Self-Adapting Filter," Information and Control, 5, 3, Sep 1962, pp. 185-194.

14. M. J. Hinich, "Large Sample Estimation of an Unknown Discrete Waveform Which Is Randomly Repeating in Gaussian Noise," TR No. 93, Dept. of Statistics, Stanford University, Stanford, Calif., Dec 1963.
15. D. J. Braverman, "Machine Learning and Automatic Pattern Recognition," TR No. 2003-1, Stanford Electronics Laboratories, Stanford, Calif., Feb 1961.
16. N. Abramson and D. Braverman, "Learning To Recognize Patterns in a Random Environment," Rept. SEL-62-071 (TR No. 2003-5), Stanford Electronics Laboratories, Stanford, Calif., May 1962. Also in IRE Trans. on Information Theory, IT-8, Sep 1962, pp. 58-63.
17. D. G. Keehn, "Learning the Mean Vector and Covariance Matrix of Gaussian Signals in Pattern Recognition," Rept. SEL-62-155 (TR No. 2003-6), Stanford Electronics Laboratories, Stanford, Calif., Feb 1963.
18. J. D. Spragins, Jr., "Reproducing Distributions for Machine Learning," Rept. SEL-63-099 (TR No. 6103-7), Stanford Electronics Laboratories, Stanford, Calif., Nov 1963.
19. R. F. Daly, "Adaptive Binary Detectors," TR No. 2003-2, Contract Nonr 225(24), Stanford Electronics Laboratories, Stanford, Calif., 26 Jun 1961.
20. R. F. Daly, "The Adaptive Binary-Detection Problem on the Real Line," Rept. SEL-62-030 (TR No. 2003-3), Stanford Electronics Laboratories, Stanford, Calif., Feb 1962.
21. R. L. Stratanovich, "Sampling of a Variable Frequency Signal from the Background Noise," Radiotekhnika, Feb 1962, pp. 171-178.
22. C. W. Helstrom, Statistical Theory of Signal Detection, Pergamon Press, Ltd., London, 1960.
23. L. A. Wainstein and V. D. Zubakov, Extraction of Signals from Noise, Translated by R. A. Silverman, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1962.
24. S. Fralick, et al, "On Learning To Detect Narrowband Signals of Unknown Frequency," Tech. Memo. No. 63 MX 13, Sylvania Electronic Systems-West, Mt. View, Calif., Jun 1963.
25. W. B. Davenport and W. L. Root, An Introduction to the Theory of Random Signals and Noise, McGraw-Hill Book Co., Inc., New York, 1958, p. 93.

26. N. Abramson and J. Farison, "On Statistical Communication Theory," Rept. SEL-62-078 (TR No. 2005-1), Stanford Electronics Laboratories, Stanford, Calif., Aug 1962.
27. C. Shannon, "Bounds on the Tails of Martingales and Related Questions," (Unpublished paper), Bell Telephone Laboratories, Murray Hill, N.J.
28. G. L. Turin, "Error Probabilities for Binary Symmetric Ideal Reception through Nonselective Slow Fading and Noise," Proc. IRE, 46, Sep 1958, pp. 1603-1619.
29. W. R. Kincheloe, Jr., "The Measurement of Frequency with Scanning Spectrum Analyzers," Rept. SEL-62-098 (TR No. 557-2), Stanford Electronics Laboratories, Stanford, Calif., Oct 1962.
30. D. Wilson, S. Fralick, and J. Knox-Seith, "Spectrum Analyzers and Sweeping Periodogram Calculators," Tech. Memo. No. 63 MX 15, Sylvania Electronic Systems-West, Mt. View, Calif., Jun 1963.
31. E. B. Dynkin, "Necessary and Sufficient Statistics for a Family of Probability Distributions," Selected Translations in Mathematical Statistics and Probability, 1, 1961, pp. 17-40.
32. T. Grottenberg, private communication.
33. D. Blackwell and M. A. Gershick, Theory of Games and Statistical Decisions, John Wiley & Sons, New York, 1954.
34. J. Walsh, Handbook of Nonparametric Statistics, John Wiley & Sons, New York, 1962.
35. J. L. Doob, Stochastic Processes, D. Van Nostrand, Co., Inc., Princeton, N.J., 1962.
36. M. Loéve, Probability Theory, 2nd Ed., D. Van Nostrand Co., Inc., Princeton, N.J., 1960.

SYSTEMS THEORY LABORATORY
DISTRIBUTION LIST
August 1964

GOVERNMENT

USAEI

1 Attn: Dr. M. Jacobs,
AMSEL-RD/SL-PF

Procurement Data Division
USAS Equipment Support Agency
Ft. Monmouth, N.J.

1 Attn: Mr. M. Rosenfeld
Commanding General, USAEI
Ft. Monmouth, N.J.

5 AMSEL-RD/SL-SC, Bldg. 42
1 TDC, Evans Signal Lab Area

Commanding Officer, ERDL
Ft. Belvoir, Va.
1 Attn: Tech. Doc. Ctr.

Commanding Officer:
Frankford Arsenal
Bridge and Talony St.
Philadelphia 37, Pa.
1 Attn: Library Br., 0270, Bldg. 40

Ballistics Research Lab
Aberdeen Proving Ground, Md.
2 Attn: V. W. Richard, BML

Chief of Naval Research
Navy Dept.
Washington 25, D.C.
2 Attn: Code 427
1 Code 420
1 Code 463

U.S. Army Electr. Labs.
Mt. View Office
P.O. Box 205
1 Mt. View, Calif.

Commanding Officer
ONR Branch Office
1000 Geary St.
1 San Francisco 9, Calif.

Chief Scientist
ONR Branch Office
1030 E. Green St.
1 Pasadena, Calif.

Office of Naval Research
Branch Office Chicago
230 N. Michigan Ave.
1 Chicago 1, Ill.

Commanding Officer
ONR Branch Office
207 W. 24th St.
New York 11, N.Y.
1 Attn: Dr. I. Rowe

U.S. Naval Applied Science Lab.
Tech. Library
Bldg. 291, Code 9832
Naval Base
1 Brooklyn, N.Y. 11251

Chief Bureau of Ships
Navy Dept.
Washington 25, D.C.
1 Attn: Code 691A1
1 Code 686
1 Code 607 NTDS
1 Code 687D
1 Code 732, A. E. Smith
1 Code 681A

Officer in Charge, ONR
Navy 100 Bx. 39, Fleet P.O.
16 New York N.Y.

U.S. Naval Research Lab
Washington 25, D.C.

6 Attn: Code 2000
1 5240
1 5430
1 5200
1 5300
1 5400
1 5266, G. Abraham
1 2027
1 5260
1 6430

Chief, Bureau of Naval Weapons
Navy Dept.
Washington 25, D.C.

1 Attn: RAAV-6
1 RUUC-1
2 RREN-3
1 RAAV-44

Chief of Naval Operations
Navy Dept.
Washington 25, D.C.
1 Attn: Code Op 945Y

Director, Naval Electronics Lab
1 San Diego 52, Calif.

USN Post Graduate School
1 Monterey, Calif.
1 Attn: Tech. Reports Librarian
1 Prof. Gray, Electronics Dept.
1 Dr. H. Titus

Weapons Systems Test Div.
Naval Air Test Center
Patuxent River, Md.
1 Attn: Library

U.S. Naval Weapons Lab
Dahlgren, Va.
1 Attn: Tech. Library

Naval Ordnance Lab.
Corona, California
1 Attn: Library
1 H. N. Wieder, 423

Commanding Officer (ADL)
USN Air Dev. Ctr.
1 Johnsville, Pa. 18974

Commander
USN Missile Center
Pt. Mugu, Calif.
1 Attn: NO3022

Commanding Officer
U.S. Army Research Office
Box CM, Duke Station
Durham, N.C.

3 Attn: CRD-AA-IP
Commanding General
U.S. Army Materiel Command
Washington 25, D.C.
1 Attn: AMCRD-DE-K
1 AMCRD-RS-PE-E

Department of the Army
Office, Chief of Res. and Dev.
The Pentagon
Washington 25, D.C.
1 Attn: Research Support Div.,
Rm. 3D442

Office of the Chief of Engineers
Dept. of the Army
Washington 25, D.C.
1 Attn: Chief, Library Br.
Hq., U.S. Air Force
Washington 25, D.C. 20330
1 Attn: AFRSTE

Aeronautical Systems Div.
Wright-Patterson AFB, Ohio

1 Attn: Lt. Col. L. M. Butsch, Jr.
ASRNE-2
1 ASRNE-2, D. R. Moore
1 ASRNR-32
1 ASRNE-1, Electronic Res. Br.
Elec. Tech. Lab
1 ASRNCF-2, Electromagnetic
and Comm. Lab
6 ASRNE-32

Systems Engineering Group (RTD,
Wright-Patterson AFB, Ohio 45433
1 Attn: SEPIR

Commandant
AF Institute of Technology
Wright-Patterson AFB, Ohio
1 Attn: AFIT (Library)

Executive Director
AF Office of Scientific Res.
Washington 25, D.C.
1 Attn: SREE

AFWL (WLL)
2 Kirtland AFB, New Mexico

Director
Air University Library
Maxwell AFB, Ala.

1 Attn: CR-4582
Commander, AF Cambridge Res. Labs
ARDC, L. G. Hanscom Field
Bedford, Mass.
1 Attn: CRTOTT-2, Electronics

Hqs., AF Systems Command
Andrews AFB
Washington 25, D.C.
1 Attn: SCTAE

Asst. Secy. of Defense (R and D/
R and D. Board, Dept. of Defense
Washington 25, D.C.
1 Attn: Tech. Library

Office of Director of Defense
Dept. of Defense
Washington 25, D.C.
1 Attn: Research and Engineering

Institute for Defense Analyses
1666 Connecticut Ave.
Washington 9, D.C.
1 Attn: W. E. Bradley

Defense Communications Agency
Dept. of Defense
Washington 25, D.C.
1 Attn: Code 121A, Tech. Library

Advisory Group on Electron Devices
346 Broadway, 8th Floor East
New York 13, N.Y.
2 Attn: H. Sullivan

Advisory Group on Reliability of
Electronic Equipment
Office Asst. Secy. of Defense
The Pentagon
1 Washington 25, D.C.

Commanding Officer
Diamond Ordnance Fuze Labs
Washington 25, D.C.
2 Attn: ORDTL 930, Dr. R.T. Young
Diamond Ordnance Fuze Lab.
U.S. Ordnance Corps
Washington 25, D.C.
1 Attn: ORDTL-450-638
Mr. R. M. Comyn

U. S. Dept. of Commerce
National Bureau of Standards
Boulder Labs
Central Radio Propagation Lab.
1 Boulder, Colorado
2 Attn: Miss J.V. Lincoln, Chief
RWSS

NSF, Engineering Section
1 Washington, D.C.

Information Retrieval Section
Federal Aviation Agency
Washington, D.C.
1 Attn: MS-112, Library Branch

DDC
Cameron Station
Alexandria 4, Va.
30 Attn: TISIA

U.S. Coast Guard
1300 E. Street, N.W.
Washington 25, D.C.
1 Attn: EEE Station 5-5

Office of Technical Services
Dept. of Commerce
1 Washington 25, D.C.

Director
National Security Agency
Fort George G. Meade, Md.
1 Attn: R42

NASA, Goddard Space Flight Center
Greenbelt, Md.
1 Attn: Code 611, Dr. G. H. Ludwig
1 Chief, Data Systems Divisions

NASA
Office of Adv. Res. and Tech.
Federal Office Bldg. 10-B
600 Independence Ave.
Washington, D.C.
1 Attn: Mr. Paul Johnson
Chief, U.S. Army Security Agency
Arlington Hall Station
2 Arlington 12, Virginia

SCHOOLS

*U of Aberdeen
Dept. of Natural Philosophy
Marischal College
Aberdeen, Scotland
1 Attn: Mr. R.V. Jones

U of Arizona
EE Dept.
Tucson, Ariz.
1 Attn: R.L. Walker
1 D.J. Hamilton

*U of British Columbia
Vancouver 8, Canada
1 Attn: Dr. A.C. Soudack

California Institute of Technology
Pasadena, Calif.
1 Attn: Prof. R.W. Gould
1 D. Braverman, EE Dept.

California Institute of Technology
4800 Oak Grove Drive
Pasadena 3, Calif.
1 Attn: Library, Jet Propulsion Lab.

U. of California
Berkeley 4, Calif.
1 Attn: Prof. R.M. Saunders, EE Dept.
Dr. R.K. Wakerling,
Radiation Lab. Info. Div.
Bldg. 30, Rm. 101

U of California
Los Angeles 24, Calif.
1 Attn: C.T. Leondes, Prof. of
Engineering, Engineering Dept.
1 R.S. Elliott, Electromagnetics
Div., College of Engineering

U of California, San Diego
School of Science and Engineering
La Jolla, Calif.
1 Attn: Physics Dept.

Carnegie Institute of Technology
Schenley Park
Pittsburg 13, Pa.
1 Attn: Dr. E.M. Williams, EE Dept.

Case Institute of Technology
Engineering Design Center
Cleveland 6, Ohio
1 Attn: Dr. J. B. Newick, Director

Cornell U
Cognitive Systems Research Program
Ithaca, N.Y.
1 Attn: F. Rosenblatt, Hollister Hall

Thayer School of Engr.
Dartmouth College
Hanover, New Hampshire
1 Attn: John W. Strohbehn
Asst. Professor

Drexel Institute of Technology
Philadelphia 4, Pa.
1 Attn: F. B. Haynes, EE Dept.

U of Florida
Engineering Bldg., Rm. 336
Gainesville, Fla.
1 Attn: M.J. Wiggins, EE Dept.

Georgia Institute of Technology
Atlanta 13, Ga.
1 Attn: Mrs. J.H. Crosland, Librarian
1 F. Dixon, Engr. Experiment
Station

Harvard U
Pierce Hall
Cambridge 38, Mass.
1 Attn: Dean H. Brooks, Div. of Engr.
and Applied Physics, Rm. 217
2 E. Farkas, Librarian, Rm. 303A,
Tech. Reports Collection

U of Hawaii
Honolulu 14, Hawaii
1 Attn: Asst. Prof. K. Najita, EE Dept.

U of Illinois
Urbana, Ill.
1 Attn: P.D. Coleman, EE Res. Lab.
1 W. Perkins, EE Res. Lab.
1 A. Albert, Tech. Ed., EE Res. Lab.
1 Library Serials Dept.
1 Prof. D. Alpert, Coordinated
Sci. Lab.

State University of Iowa
Dept. of Electrical Engineering
Iowa City, Iowa
1 Attn: Prof. Donald L. Epley

*Instituto de Pesquisas da Marinha
Ministerio da Marinha
Rio de Janeiro
Estado da Guanabara, Brazil
1 Attn: Roberto B. da Costa

Johns Hopkins U
Charles and 34th St.
Baltimore 18, Md.
1 Attn: Librarian Carlisle Barton Lab.

Johns Hopkins U
8621 Georgia Ave.
Silver Spring, Md.
1 Attn: N. H. Chokay
1 Mr. A.W. Nagy, Applied
Physics Lab.

Linfield Research Institute
McMinnville, Ore.
1 Attn: G. N. Hickok, Director

Marquette University
College of Engineering
1515 W. Wisconsin Ave.
Milwaukee 3, Wis.
1 Attn: A.C. Moeller, EE Dept.

M I T
Cambridge 39, Mass.
1 Attn: Res. Lab. of Elec., Doc. Rm.
1 Miss A. Sils, Libn. Rm. 4-244,
LIR
1 Mr. J.E. Ward, Elec. Sys. Lab.

M I T
Lincoln Laboratory
P.O. Box 73
1 Attn: Lexington 73, Mass.
1 Navy Representative
1 Dr. W.I. Wells
1 Kenneth L. Jordan, Jr.

U of Michigan
Ann Arbor, Mich.
1 Attn: Dir., Cooley Elec. Labs.,
N. Campus
1 Dr. J.E. Rowe, Elec. Phys. Lab.
1 Comm. Sci. Lab., 180 Frieze Bldg

U of Michigan
Institute of Science and Technology
P.O. Box 618
Ann Arbor, Mich.
1 Attn: Tech. Documents Service
1 W. Wolfe--IRIA--

U of Minnesota
Institute of Technology
Minneapolis 14, Minn.
1 Attn: Prof. A. Van der Ziel,
EE Dept.

U of Nevada
College of Engineering
Reno, Nevada
1 Attn: Dr. R.A. Manhart, EE Dept.

Northeastern U
The Dodge Library
Boston 15, Mass.
1 Attn: Joyce E. Iander, Librarian

Northwestern U
2422 Oakton St.
Evanston, Ill.
1 Attn: W.S. Toth Aerial
Measurements Lab.

U of Notre Dame
South Bend, Ind.
1 Attn: E. Henry, EE Dept.

Ohio State U
2024 Niel Ave.
Columbus 10, Ohio
1 Attn: Prof. E.M. Boone, EE Dept.

Oregon State U
Corvallis, Ore.
1 Attn: H.J. Oorthuys, EE Dept

Polytechnic Institute
333 Jay St.
Brooklyn, N.Y.
1 Attn: L. Shaw, EE Dept.

* No AF or Classified Reports

Polytechnic Institute of Brooklyn
Graduate Center, Route 110
Farminigdale, N.Y.
1 Attn: Librarian

Purdue U
Lafayette, Ind.
1 Attn: Library, EE Dept.

Rensselaer Polytechnic Institute
School of Engineering
Troy, N.Y.
1 Attn: Library, Serials Dept.
1 Kenneth E. Mortenson

*U of Saskatchewan
College of Engineering
Saskatoon, Canada
1 Attn: Prof. R.E. Ludwig

Syracuse U
Syracuse 10, N.Y.
1 Attn: EE Dept.

*Uppsala U
Institute of Physics
Uppsala, Sweden
1 Attn: Dr. P. A. Tove

U of Toledo
Dept. of Electr. Engr.
Toledo 6, Ohio
1 Attn: James B. Farison
Asst. Prof.

U of Utah
Salt Lake City, Utah
1 Attn: R.W. Grow, EE Dept.

U of Virginia
Charlottesville, Va.
1 Attn: J.C. Wyllie, Alderman
Library

U of Washington
Seattle 5, Wash.
1 Attn: A. E. Harrison, EE Dept.

Worcester Polytechnic Inst.
Worcester, Mass.
1 Attn: Dr. H.M. Newell

Yale U
New Haven, Conn.
1 Attn: Sloane Physics Lab.
1 EE Dept.
1 Dunham Lab., Engr. Library

INDUSTRIES

Avco Corp.
Res. Lab.
2385 Revere Beach Parkway
Everett 49, Mass.
1 Attn: Dr. Gordon Abell

Argonne National Lab.
9700 South Cass
Argonne, Ill.
1 Attn: Dr. O.C. Simpson

Admiral Corp.
3800 Cortland St.
Chicago 47, Ill.
1 Attn: E.N. Robertson, Librarian

Airborne Instruments Lab.
Comac Road
Deer Park, Long Island N.Y.
1 Attn: J. Dyer, Vice-Pres. and
Tech. Dir.

Asperex Corp.
230 Duffy Ave.
Hicksville, Long Island, N.Y.
1 Attn: Prot. Engineer, S. Barbasso

*No AF or Classified Reports.

Autonetics
Div. of North American Aviation, Inc.
9150 E. Imperial Highway
Downey, Calif.
1 Attn: Tech. Library 3040-3

Bell Telephone Labs.
Murray Hill Lab.
Murray Hill, N.J.
1 Attn: Dr. J.R. Pierce
1 Dr. S. Darlington
1 Mr. A.J. Grossman

Bell Telephone Labs., Inc.
Technical Information Library
Whippany, N.J.
1 Attn: Tech. Repts. Librn.,
Whippany Lab.

The Boeing Company
Mail Stop MS-1331-ORG. 1-8000
Seattle 24, Washington
1 Attn: Dr. Ervin J. Nalos

*Central Electronics Engineering
Research Institute
Pilani, Rajasthan, India
1 Attn: Om P. Gandhi - Via: ONR/London

Columbia Radiation Lab.
538 West 120th St.
New York, New York
1 Attn: Convair - San Diego
Div. of General Dynamics Corp.
San Diego 12, California
1 Attn: Engineering Library

Cook Research Labs
6401 W. Oakton St.
1 Attn: Morton Grove, Ill.

Cornell Aeronautical Labs., Inc.
4455 Genessee
Buffalo 21, N.Y.
1 Attn: Library

Eitel-McCullough, Inc.
301 Industrial Way
San Carlos, Calif.
1 Attn: Research Librarian

Ewan Knight Corp.
East Natick, Mass.
1 Attn: Library

Fairchild Semiconductor Corp.
4001 Junipero Serra Blvd.
Palo Alto, Calif.
1 Attn: Dr. V. H. Grinich

General Electric Co.
Defense Electronics Div., LMED
Cornell University, Ithaca, N.Y.
1 Attn: Library
Via: Commander, ASD W-P, Ohio,
ASRNGW D.E. Lewis

General Electric TWT Products Sec.
601 California Ave.
Palo Alto, Calif.
1 Attn: Tech. Library, C.G. Lob

General Electric Co. Res. Lab.
P.O. Box 1088
Schnectady, N.Y.
1 Attn: Dr. P.M. Lewis
1 R.L. Shuey, Mgr. Info.
Studies Sec.

General Electric Co.
Electronics Park
Bldg. 3, Rm 143-1
Syracuse, N.Y.
1 Attn: Doc Library, Y Burke

Gilfillan Brothers
1815 Venice Blvd.
Los Angeles, Calif.
1 Attn: Engr. Library

The Hallicrafters Co.
5th and Kostner Ave.
1 Attn: Chicago 24, Ill.

Hewlett-Packard Co.
1501 Page Mill Road
1 Attn: Palo Alto, Calif.

Hughes Aircraft
Malibu Beach, Calif.
1 Attn: Mr. Iams

Hughes Aircraft Co.
Florence at Teale St.
Culver City, Calif.
1 Attn: Tech. Doc. Cen., Bldg. 8,
Rm. C2048

Hughes Aircraft Co.
P.O. Box 278
Newport Beach, Calif.
1 Attn: Library, Semiconductor Div.

IBM, Box 390 Boardman Road
Poughkeepsie, N.Y.
1 Attn: J.C. Logue, Data Systems Div.

IBM Poughkeepsie, N.Y.
1 Attn: Product Dev. Lab.,
E.M. Davis

IBM ASD and Research Library
Monterey and Cortle Roads
San Jose, Calif.
1 Attn: Miss M. Griffin, Bldg. 025

ITT Federal Labs.
500 Washington Ave.
Nutley 10, N.J.
1 Attn: Mr. E. Mount, Librarian

Laboratory for Electronics, Inc.
1075 Commonwealth Ave.
Boston 15, Mass.
1 Attn: Library

LEL, Inc.
75 Akron St.
Copiasque, Long Island, N.Y.
1 Attn: Mr. R.S. Mautner

Lenkurt Electric Co.
San Carlos, Calif.
1 Attn: M.L. Waller, Librarian

Librascope
Div. of General Precision, Inc.
808 Western Ave.
Glendale 1, Calif.
1 Attn: Engr. Library

Lockheed Missiles and Space Div.
P.O. Box 504, Bldg. 524
Sunnyvale, Calif.
1 Attn: Dr. W.M. Harris, Dept. 65-70
1 G. W. Price, Dept. 67-33

Melpar, Inc.
3000 Arlington Blvd.
Falls Church, Va.
1 Attn: Librarian

Microwave Associates, Inc.
Northwest Industrial Park
Burlington, Mass.
1 Attn: K. Mortenson
1 Librarian

Microwave Electronics Corp.
4061 Transport St.
Palo Alto, Calif.
1 Attn: S.F. Kiesel
M.C. Long

Minneapolis-Honeywell Regulator Co.
1177 Blue Heron Blvd
Riviera Beach, Fla
1 Attn: Semiconductor Products Library

The Mitre Corp.
Bedford, Mass.
1 Attn: Library

Monsanto Research Corp. Station B, Box 8 Dayton 7, Ohio 1 Attn: Mrs. D. Crabtree	Sperry Rand Corp. Sperry Electron Tube Div. Gainesville, Fla. 1 Attn: Librarian
Monsanto Chemical Co. 800 N. Linbergh Blvd. St. Louis 66, Mo. 1 Attn: Mr. E. Orban, Mgr. Inorganic Dev.	Sperry Gyroscope Co. Div. of Sperry Rand Corp. Great Neck, N.Y. 1 Attn: L. Swern (MS3T105)
*Dir., National Physical Lab. Hillside Road New Delhi 12, India 1 Attn: S.C. Sharma - Via: ONR London	Sperry Gyroscope Co. Engineering Library Mail Station F-7 Great Neck, Long Island, N.Y. 1 Attn: K. Barney, Engr. Dept. Head
*Northern Electric Co., Ltd. Research and Development Labs. P.O. Box 3511, Station "C" Ottawa, Ontario, Canada 1 Attn: J.F. Tatlock	Sperry Microwave Electronics Clearwater, Fla. 1 Attn: J.E. Pippin, Res. Sec. Head
Northronics Palos Verdes Research Park 6101 Crest Road Palos Verdes Estates, Calif. 1 Attn: Tech. Info. Center	Sylvania Electric Products, Inc. 500 Evelyn Ave. 1 Mt. View Calif. 1 Attn: Mr. E.O. Ammann
Pacific Semiconductors, Inc. 14520 So. Aviation Blvd. Lawndale, Calif. 1 Attn: H.Q. North	Sylvania Electronics Systems 100 First St. Waltham 54, Mass. 1 Attn: Librarian, Waltham Labs. 1 Mr. E.E. Hollis
Philco Corp. Tech. Rep. Division P.O. Box 4730 Philadelphia 34, Pa. 1 Attn: F.R. Sherman, Mgr. Editor	Technical Research Group Route No. 117 1 Melville, New York 11749
Philco Corp. Jolly and Union Meeting Roads Blue Bell, Pa. 1 Attn: C.T. McCoy 1 Dr. J. R. Feldmeier	Texas Instruments, Inc. P.O. Box 6015 Dallas 22, Texas 1 Attn: M.E. Chun, Apparatus Div.
Polarad Electronics Corp. 43-20 Thirty-Fourth St. Long Island City 1, N.Y. 1 Attn: A.H. Sonnenschein	Texas Instruments, Inc. P.O. Box 5012 Dallas, Texas 75222 1 Attn: Tech. Repts. Service, MS-65 2 Semi-Conductor Components Library
Radio Corp. of America RCA Labs., David Sarnoff Res. Cen. Princeton, N.J. 2 Attn: Dr. J. Sklansky	Texas Instruments, Inc. Corporate Research and Engineering Technical Reports Service P.O. Box 5474 1 Dallas 22, Texas
RCA Labs., Princeton, N.J. 1 Attn: H. Johnson	Tektronix, Inc. P.O. Box 500 Beaverton, Ore. 4 Attn: Dr. J.F. DeLord, Dir. of Research
RCA, Missile Elec. and Controls Dept. Woburn, Mass. 1 Attn: Library	Varian Associates 611 Hansen Way Palo Alto, Calif. 1 Attn: Tech. Library
The Rand Corp. 1700 Main St. Santa Monica, Calif. 1 Attn: Helen J. Waldron, Librarian	Weitermann Electronics 4549 North 38th St. 1 Milwaukee 9, Wisconsin
Raytheon Manufacturing Co. Microwave and Power Tube Div. Burlington, Mass. 1 Attn: Librarian, Spencer Lab.	Westinghouse Electric Corp. Friendship International Airport Box 746, Baltimore 3, Md. 1 Attn: G.R. Kilgore, Mgr. Appl. Res. Dept. Baltimore Lab.
Raytheon Manufacturing Co. Res. Div., 28 Seyon St. Waltham, Mass. 1 Attn: Dr. H. Stutz 1 Mrs. X. Bennett, Librarian 1 Research Div. Library	Westinghouse Electric Corp. 3 Gateway Center Pittsburgh 22, Pa. 1 Attn: Dr. G.C. Sziklai
Roger White Electron Devices, Inc. Tall Oaks Road 1 Laurel Hedges, Stamford, Conn.	Westinghouse Electric Corp. P.O. Box 284 Elmira, N.Y. 1 Attn: S.S. King
Sandia Corp. Sandia Base, Albuquerque, N.M. 1 Attn: Mrs. B.R. Allen, Librarian	Zenith Radio Corp. 6001 Dickens Ave. Chicago 39, Ill. 1 Attn: J. Markin

*No At or Classified Reports

UNCLASSIFIED

Security Classification

DOCUMENT CONTROL DATA - R&D		
(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)		
1. ORIGINATING ACTIVITY (Corporate author) Stanford Electronics Laboratories Stanford University, Stanford, California		2a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED
		2b. GROUP
3. REPORT TITLE Learning to Recognize Patterns without a Teacher		
4. DESCRIPTIVE NOTES (Type of report and inclusive dates) Technical Report		
5. AUTHOR(S) (Last name, first name, initial) Stanley C. Fralick		
6. REPORT DATE March 1965	7a. TOTAL NO. OF PAGES 116	7b. NO. OF REFS 36
8a. CONTRACT OR GRANT NO. ONR Contract Nonr-225(83)	8c. ORIGINATOR'S REPORT NUMBER(S) Technical Report No. 6103-10 SU-SEL-65-011	
8b. PROJECT NO.		
9.	9d. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)	
10. AVAILABILITY/LIMITATION NOTICES Foreign announcement and dissemination by DDC limited.		
11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY Jointly supported by U.S. Army Signal Corps, U.S. Air Force, and U.S. Navy	
13. ABSTRACT The techniques of decision theory are applied to the problem of constructing machines that improve their ability to recognize patterns by extracting pertinent information from a previously unclassified sequence of observations; such machines are said to learn without a teacher. A general system solution is obtained which includes the solutions to the problems of learning without a teacher, learning with a teacher, and no learning. The solution has been extended to include problems in which the unknown parameter is time varying, as well as problems in which the probabilities of occurrence of the classes are unknown a priori and must be learned. The resulting systems are shown to be stable and to have performance which converges to the performance of systems which have a priori knowledge of the unknown parameters being learned. It has been demonstrated that for most cases either the optimum system, or a sub-optimum system which performs within an arbitrarily small tolerance of the optimum system, is realizable in the sense that it requires a finite memory. The techniques of this paper are applied to examples of learning problems in the communications, radar, and electromagnetic reconnaissance fields.		

UNCLASSIFIED

Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
LEARNING MACHINES PATTERN RECOGNITION SIGNAL DETECTION SYSTEMS SYNTHESIS						

INSTRUCTIONS

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

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

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

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

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

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

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

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

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

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

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

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

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

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

imposed by security classification, using standard statements such as:

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

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

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

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

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

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

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

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