ERROR PROBABILITY IN UNSUPERVISED DEPENDENT PATTERN CLASSIFICATION

Thesis for the Degree of Ph. D.
MICHIGAN STATE UNIVERSITY
JOHN JAMES FORSYTH
1971

This is to certify that the

thesis entitled

ERROR PROBABILITY IN UNSUPERVISED DEPENDENT
PATTERN CLASSIFICATION

presented by

John James Forsyth

has been accepted towards fulfillment
of the requirements for

Ph.D. degree in Electrical Engineering & Systems Science

Richard C. Dubes
Major professor

Date Aug. 30, 1971

O-7639

ABSTRACT

ERROR PROBABILITY IN UNSUPERVISED DEPENDENT PATTERN CLASSIFICATION

By

John James Forsyth

This thesis investigates a communications problem within the framework of unsupervised multi-category pattern recognition. Consider a digital data communications system in which a source randomly selects symbols from an alphabet, encodes the symbols as a string of digits (the code being of fixed but unknown length) and transmits the resulting digital data over a channel. A symbol synchronization problem ensues when a receiver locks on to the signal at a time other than when the first digit of a coded symbol arrives. When the receiver does not know the length of the individual symbol codes being received, and when special synchronizing pulses are not present as a guide, then the receiver is faced with processing patterns which exhibit statistical dependence. The term "synchronization" as used here refers to the problem of establishing the starting point of each symbol code in the data, which in turn requires determination of the code length.

The syn

in unsupervis

solutions to

both through

stochastic ap

havior of the

whether the s

an independen

decision proc

Since the dec

possible prob

A detai

decision proc

through a lim

probability d

Another uses

of two measu

butions is ex

coefficient

ational dist

distance mea

John James Forsyth

The synchronization problem is treated as a problem in unsupervised multi-category pattern recognition. Solutions to the synchronization problem are developed both through the Bayes decision process and through a stochastic approximation algorithm. The convergent behavior of these solutions is proved. It is shown that whether the source generating the codes is governed by an independent or a Markov random process the Bayes decision process for the synchronization problem converges. Since the decision procedures use no training data, the possible probability models for the source must be known.

A detailed study of error bounds for the Bayes decision process is presented. One bound is obtained through a limiting process which examines the asymptotic probability of error of a suboptimum decision process. Another uses information theoretic concepts. The roles of two measures of distance between probability distributions is examined; those measures are the Bhattacharyya coefficient (Hellinger integral) and the Kolmogorov variational distance. Error bounds based directly on those distance measures are exhibited.

ERROR PROBABILITY IN UNSUPERVISED DEPENDENT PATTERN CLASSIFICATION

By

John James Forsyth

A THESIS

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Electrical Engineering and
Systems Science

1971

To my wife and children

Thank
for his tho
course of t
of the writ
for his hel
Professors
to particul
the efforts
G. L. Park
improvement

ACKNOWLEDGMENTS

Thanks are in order to Professor R. C. Dubes both for his thoughtful guidance and assistance in shaping the course of this research, and for his painstaking review of the writing. Thanks also go to Professor R. J. Reid for his help in defining the motivating problem, and to Professors Reid and R. Staudte for stimulating solutions to particular problems which were encountered. Finally, the efforts of Professors M. G. Keeney, C. V. Page, and G. L. Park in reviewing the thesis and suggesting helpful improvements are acknowledged.

TABLE OF CONTENTS

Table

LIST OF TABLES

Figure

LIST OF FIGURES

Notation

A

A_k

B

B_k

$\mathrm{cov}_i(V)$

$d(X^{2n+1})$

$d_B(X^k)$

d_{CB}

$d_k(X_k)$

D

LIST OF NOTATION

Notation	Meaning	Page Reference
A	$\lim_{n\to\infty} \frac{1}{n} \sum_{h=1}^{n} A_{k+h}$	52
A_k	$2\text{cov}_1\{U_k, U_{k+1}\} + \text{var}_1\{U_{k+1}\}$	52
B	$\lim_{n\to\infty} \frac{1}{n} \sum_{h=1}^{n} B_{k+h}$	56
B_k	$2\text{cov}_2\{V_k, V_{k+1}\} + \text{var}_2\{V_{k+1}\}$	56
$\text{cov}_i(V)$	Covariance of V under class i	52
$d(X^{2n+1})$	Majority decision function	47
$d_B(X^k)$	Bayes decision function	9
d_{CB}	Compound Bayes majority decision function	65
$d_k(X_k)$	Decision function used for k-th pattern for majority decision function	47
D	Majority decision function--same as $d(X^{2n+1})$	47

Notation

$E(V|\Theta = t_i)$

$E(V|\Omega_i)$

$E_i(V)$

$E_G(V)$

$f(V|X^k)$

$f_i(V)$

g_{i_o}

$g_{i,n}$

$\bar{g}_{i,n}$

$G = \{g_1, \ldots$

$H(\cdot)$

H_j

J

Notation	Meaning	Page Reference
$E(V\|\theta = t_i)$	Conditional expectation of V under class i	52
$E(V\|\Omega_i)$	Same as $E(V\|\theta = t_i)$	74
$E_i(V)$	Same as $E(V\|\theta = t_i)$	11
$E_G(V)$	Expectation of V with respect to the mixture G	33
$f(V\|X^k)$	Sample conditional probability density function for V	37
$f_i(V)$	Conditional probability density function for V under class i	11
g_{i_o}	Element of G that equals one	30
$g_{i,n}$	Estimator for g_i based on n observations	33
$\bar{g}_{i,n}$	Non-negative estimator for g_i derived from $g_{i,n}$	33
$G = \{g_1,\ldots,g_m\}$	Probability distribution for λ used as a finite mixing distribution	30
$H(\cdot)$	Shannon's entropy	69
H_j	A vector space	32
J	Divergence measure relative to two probability density functions	11

Notation

ℓ

ℓ_o

$\hat{\ell}_{kB}$

$\hat{\ell}_{kM}$

L

m_i

n

p_i

$P(V|W)$

$P_e(V)$

$\bar{P}_m(V)$

$P_i(V)$

$P_i(V|W)$

$\underline{P}_i(V)$

Notation	Meaning	Page Reference
ℓ	Index of sources	19
ℓ_o	Correct decision about ℓ	26
$\hat{\ell}_{kB}$	Bayes estimator for ℓ	25
$\hat{\ell}_{kM}$	Maximum likelihood extimator for ℓ	26
L	Total number of sources	19
m_ℓ	Symbol code length for source ℓ	19
n	Number of digits in a pattern	20
p_i	Prior probability that class i is active	46
$P(V\|W)$	Conditional Probability mass function for V given W	23
$P_e(V)$	Probability of error for decision function V	49,66
$\bar{P}_m(V)$	Empirical probability mass function for V	28
$P_i(V)$	Probability mass function for V under class i	46
$P_i(V\|W)$	Conditional probability mass function for V given W under class i	9
$\underline{P}_i(V)$	Vector of probability mass function for V under class i	32

Notation

$P_o(V)$

$r_{ij}(X_k)$

S_{ik}

t_i

T_k

$\hat{T}_{kB}$

$\hat{T}_{kM}$

T_{ok}

U_k

V_k

$var_i(V)$

X_k

X^k

Notation	Meaning	Page Reference
$P_o(V)$	Prior probability mass function for V	25
$r_{ij}(X_k)$	Factor for recursively computing $\rho_{ij}^{(k)}$	84
S_{ik}	Decision region for deciding class i at k-th observation	47
t_i	Value of θ indicating a pattern class	9
T_k	Synchronization point in k-th pattern	20
$\hat{T}_{kB}$	Bayes estimator for T_k	25
$\hat{T}_{kM}$	Maximum likelihood estimator for T_k	26
T_{ok}	Correct decision about T_k	26
U_k	Decision indicator random variable for k-th pattern-- $U_k = 1 - V_k$	48
V_k	Decision indicator random variable for k-th pattern-- $V_k = 1 - U_k$	48
$var_i(V)$	Variance of V under class i	52
X_k	Random variable for k-th pattern	9,20
X^k	The sequence of random variables $X_1,\ldots,X_k$	9,20

Notation

Z_j

$\bar{a}$

$\underline{a}$

a_i

a_{ik}

$a_{ik}(U_{k-1})$

$\gamma_k(\ell, T_k)$

δ

ε

η

θ

λ

ρ

Notation	Meaning	Page Reference
Z_j	Base 10 representation of a binary pattern	99
$\overline{\alpha}$	$\max(\alpha_1,\alpha_2)$	59
$\underline{\alpha}$	$\min(\alpha_1,\alpha_2)$	59
α_i	Probability of error under class i	58
α_{ik}	Probability of error of $d_k(X_k)$ under class i	47
$\alpha_{ik}(U_{k-1})$	Probability of error of $d_k(X_k)$ given $d_{k-1}(X_{k-1})$ under class i	50
$\gamma_k(\ell,T_k)$	Mapping from (ℓ,T_k) to λ	30
δ	One-half of upper bound on variational distance	59
ε	$\varepsilon = \frac{1}{2} - \frac{\delta}{4}$	62
η	$\eta = \overline{\alpha} - \underline{\alpha}$	59
Θ	Class indexing parameter	9
λ	Class indexing parameter mapped from (ℓ,T_k) under $\gamma_k(\ell,T_k)$	30
ρ	Bhattacharyya coefficient (Hellinger integral)	11

Notation

c_{ij}

$c_{ij}^{(k)}$

ϕ_i

$\phi(x,u,\sigma^2)$

$\hat{\mu}_h$

Notation	Meaning	Page Reference
ρ_{ij}	Interclass Bhattacharyya coefficient based on one pattern	81
$\rho_{ij}^{(k)}$	Interclass Bhattacharyya coefficient based on k patterns	82
ϕ_i	Component of $\underline{P}_i(V)$	33
$\Phi(x,u,\sigma^2)$	Integral over the right tail of the normal density function	55
Ω_h	Subset of a probability space on which $\Theta = t_h$	74

Recer

engineering

use of stat

the traditi

H-2, P-2, W

trend towar

the framewo

estimation,

formulation

assumptions

system. In

the applica

edge of pro

ticians hav

Bayes strat

tomed to th

system desc

*Alth

the Bayesia

CHAPTER I

INTRODUCTION

Recent research efforts in engineering and recent engineering practice have turned more and more toward the use of statistical models rather than, or in addition, to the traditional deterministic system models [A-7, D-2, F-5, H-2, P-2, W-1]. In particular, there exists a growing trend toward the formulation of engineering problems in the framework of statistical hypothesis testing, parameter estimation, and pattern recognition methodology. These formulations often show a willingness to make certain assumptions about the probabilistic description of the system. Indeed, many of the solution algorithms grow from the application of Bayes' rule, which requires prior knowledge of probability distributions [H-7, N-1]. Some statisticians have long eschewed any such assumptions,* but the Bayes strategy represents a logical step for those accustomed to the availability of a complete, deterministic system description. As a rule, the models describe

---

*Although for a scholarly treatise supporting the Bayesian approach see Good [G-1].

process
tistica
cations
produce
process

1.1 Com

Fu
a source

The objec
priately
and whic
Radar sys
successfu
Neyman-Pe
W-1] to c
radar app
ing noise
signal is
descriptio
the parame
receiver c
minimize t
dismissal

processes for which successive experiments produce statistically independent outcomes. However, in communications theory and other fields, processes can occur which produce statistically dependent observations. One such process is now introduced.

1.1 Communications System Models

Fundamentally, a communications system consists of a source cascaded with a channel and a receiver (Figure 1.1).

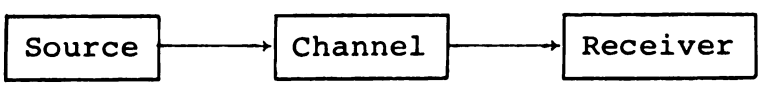


Figure 1.1 A Communication System

The objective of the system is to have the receiver appropriately process signals, which are emitted from the source, and which are subject to some modification by the channel. Radar systems provide the first historical example of the successful application of probabilistic models and of the Neyman-Pearson hypothesis testing theory [D-2, H-2, M-8, W-1] to communication systems. In the usual model of the radar application, the channel distorts the signal by adding noise to it. The receiver must determine whether a signal is present or absent. From the probabilistic descriptions of the source and channel, one can determine the parameters that define a threshold detector which the receiver can use on the output of a correlator in order to minimize the false alarm probability for a given false dismissal probability. The source and the channel can

both be r

tain eith

M-7, P-1

trates th

is more

The

nition, a

consider

sources r

its own p

So

So

Sou

Fi

from an al

and transm

The receiv

starting p

effective

study unde

both be modeled probabilistically, and the models can contain either independent or dependent random variables [M-5, M-7, P-1, S-3]. While the signal detection problem illustrates the use of probabilistic models, a modified system is more directly related to the problems of this thesis.

The fields of communication theory, pattern recognition, and statistical hypotheses testing intertwine when considering a communications system in which one of several sources may transmit signals (Figure 1.2). Each source has its own process by which it repeatedly selects symbols

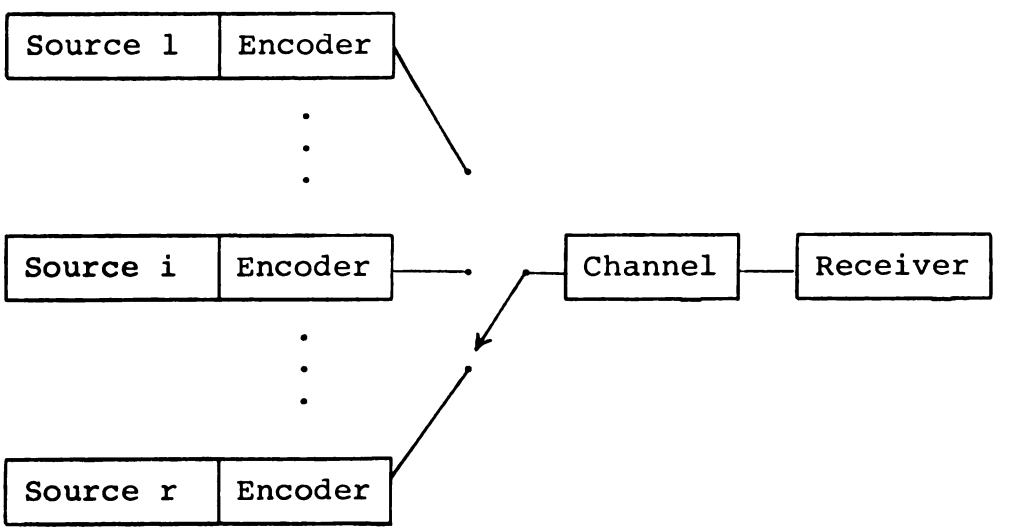


Figure 1.2 Multisource Communication System

from an alphabet, encodes the symbol into several digits, and transmits the resulting digits in a continuing stream. The receiver must know which source is active and the starting point of each symbol code in order to apply an effective decoding algorithm. The problem motivating the study undertaken in this thesis is to determine which

source

receiv

zation

starti

1.2 Sy

channel

a correl

the symb

problem

because

because

receiving

type of p

each symb

They furth

tically in

zation pro

and Stewar

proved tha

theorem [S-

One

with others

Recent work

and Murtha

source is active and the symbol code starting point. The receiver will be said to have achieved symbol synchronization when it has identified the source and determined the starting point of each symbol code which it receives.

1.2 Symbol Code Synchronization

Digital data transmitted over a communication channel often contains a symbol synchronization pulse which a correlator can process to determine the starting point of the symbol code [M-2, W-1]. A symbol synchronization problem arises when such a pulse is either not provided because of bandwidth or other design considerations, lost because of channel degradation, or simply unknown at the receiving end. Hancock and Stewart [H-1] considered this type of problem by assuming that the number of bits in each symbol code was known and fixed from symbol to symbol. They further assumed that successive symbols were statistically independent. By modeling the symbol synchronization problem as a hypothesis testing problem, Hancock and Stewart established a Bayesian decision procedure and proved that it converges in the sense of Spragins' theorem [S-6].

One should compare this type of synchronization with others receiving attention in current research. Recent works by McBride and Sage [M-3, M-4] and by Farrell and Murtha [F-1] examine the problem of extracting bit

synchro

purpose

which a

On the

synchro

symbol.

quently

that acc

able for

to devel

of sever

the type

problem

Stewart.

sources.

and succe

that alph

into a ce

used is c

mits stat

independe

uses a fi

but diffe

code leng

synchronization information from the message data. Their purpose is to determine the appropriate time interval over which a correlator should operate for signal detection. On the other hand, Harnett [H-3] takes the approach that synchronization error effectively inserts or deletes a symbol. However, except for the randomly (and infrequently) inserted and deleted symbols, Harnett assumes that accurate synchronization, or registration, is available for the rest of the symbol stream, and he proceeds to develop decoding theorems for code strings consisting of several symbols. These two views are distinct from the type of synchronization studied in this thesis.

This thesis approaches the symbol synchronization problem by generalizing the point of view of Hancock and Stewart. A receiver obtains data from one of a set of sources. The source model assumes an alphabet of symbols, and successive random experiments select symbols from that alphabet for transmission. Each symbol is encoded into a certain number of digits, and the number of digits used is called the symbol code length. Each source transmits statistically independent symbols although this independence constraint is not necessary. Each source uses a fixed symbol code length for all of its symbols, but different sources do not necessarily use the same code length.

Ex

fc

us

re

fr

wa

an

mi

tra

The

observing

first bit

the recei

that the

Example

Suppose there are two sources, one using a four binary digit (binit) code length and the other using three binits for each code word. If the receiver observes six binits, say

1 0 1 1 1 0

from the data stream, there are several possible ways these binits could be partitioned for decoding.

If the four binit source is transmitting, then any of the partitions

1 0 1 1 | 1 0 . .

. . . 1 | 0 1 1 1 | 0 . . .

. . 1 0 | 1 1 1 0

. 1 0 1 | 1 0 . .

might be appropriate. If the three binit source is transmitting, then there are three possible partitions.

1 0 1 | 1 1 0

. . 1 | 0 1 1 | 1 0 .

. 1 0 | 1 1 1 | 0 . .

The receiver does not know which source it is observing or whether the first bit it received is the first bit of a symbol code. However, it is assumed that the receiver has achieved bit synchronization, and further that the receiver knows something about the probability

distr
butio
functi
zation
proble
first
precis
must c
random

1.3 Pa

T
ature a
ables [
be said
ables [
random
extend t
examples
of joint
tation o
The conv
pendent
random va

The
literatur
ables are
weak and

distributions governing the sources--either the distributions themselves or the parametric form of their density functions. Under these assumptions, this symbol synchronization problem can be modeled as a pattern recognition problem for which the observed random variables exhibit first order dependence, or the Markov property (for the precise specification, see page 24). Consequently, one must consider decision-making processes using dependent random variables.

1.3 Pattern Recognition Methodology

There is an ample and rapidly growing body of literature about decision making based on independent random variables [D-3, D-4, L-1, P-1]. Unfortunately, the same cannot be said for decision making based on dependent random variables [H-6]. Quite often, the algorithms for independent random variables derive from some property that does not extend to dependent random variables. Two salient examples are the product rule for defining probabilities of joint independent events and the rule for the expectation of the product of uncorrelated random variables. The convenient factoring which they provide in the independent case are denied to one who deals with dependent random variables.

The most powerful general results in the statistical literature which are applicable to dependent random variables are the Central Limit Theorem [F-4, H-8], and the weak and strong laws of large numbers [F-2]. These tools

provi
of de
the c
for Pa
Stewar
consis
zation
exhibi
Bayes
the pr
ponds t
mate a
bution
the mod
nonzero
of the
puted by
consiste
the post
the prob
Conseque
lated ar
observat
argument
questions
and stora

provide the basis for describing the asymptotic behavior of decision processes. Signori [S-3] used them to show the convergence of optimum (Bayesian) decision processes for Partially Observable Markov Systems. Hancock and Stewart also used them to show the existence of a strongly consistent sequence of estimators for the symbol synchronization problem. In both works, the technique was to exhibit strongly consistent estimators in proving that the Bayes posterior distributions asymptotically assign all the probability mass to the pattern class which corresponds to the correct decision [S-5, S-6]. One can estimate a parameter value by computing the posterior distribution of the parameter and taking the mean argument of the mode as the estimate. Spragins showed that if a nonzero prior probability is assigned to the true value of the parameter, if the posterior distributions are computed by the Bayes rule and if there exists a strongly consistent sequence of estimators for the parameter, then the posterior distribution asymptotically assigns all of the probability mass to the true value of the parameter. Consequently, when the posterior distributions so calculated are premised on a sufficiently large number of observations, the parameter estimate provided by the mean argument of the mode should be reliable. This raises questions of rate of convergence, probability of error, and storage requirements which are examined in this thesis.

M

process

Given a

variabl

i.e., p

The Bay

and a p

$\theta = \{t_1$

require

distrib

P

where X^k

mass fu

the obse

d_B

P(

Here, d_B

is the i

*F

defined

values o

Much of what follows concerns the Bayes decision process, which will be briefly summarized at this point. Given a sequence of observations, X_1, X_2, ..., of a random variable governed by one of m probability distributions, i.e., pattern classes, decide which distribution is active. The Bayes decision process requires a prior distribution and a parameter, say Θ, which indicates the active class, $\Theta = \{t_1, \ldots, t_m\}$. The Bayes decision process further requires calculating the sample conditional posterior distribution of Θ by the Bayes rule:*

$$P(\Theta = t_i | X^k) = \frac{P(\Theta = t_i | X^{k-1})\ P_i(X_k | X^{k-1})}{\sum_{\Theta} P(\Theta = t_i | X^{k-1})\ P_i(X_k | X^{k-1})}$$

where $X^k = X_1, X_2, \ldots, X_k$ and $P_i(X_k | X^{k-1})$ is the i-th mass function of X_k. The decision function, which maps the observation space to the decision space, is

$d_B(X^k) = t_i$ if i is the smallest integer for which

$P(\Theta = t_i | X^k) \geq P(\Theta = t_j | X^k)$, all $j \neq i$.

Here, $d_B(X^k) = t_i$ means to decide that element t_i of Θ is the index of the pattern class being observed.

---

*For notational convenience, terms which have been defined to denote random variables will also denote the values of the random variables.

U

its wel

risk st

minimum

generat

to an e

the min

U

Bayes de

pendent,

vestigat

R-4]. C

suboptim

rule. T

optimum

behavior

rule, and

the numb

Sev

of boundi

of "dista

classes.*

likelihoo

yet been

the dista

* T

the meas

required

Use of the Bayes decision process is motivated by its well-known properties which include being the minimum risk strategy and being the strategy which leads to minimum probability of error. A standard derivation generates the Bayes decision function as the solution to an extremum problem, in which the extremum sought is the minimum of the expectation of a risk function.

Upper bounds on the probability of error for the Bayes decision process have been established for independent, continuous random variables by a number of investigators [C-10, F-3, H-5, K-2, L-2, L-3, L-4, R-3, R-4]. Chu and Chueh established bounds by defining a suboptimum decision rule called the majority decision rule. The Central Limit Theorem implied that the suboptimum procedure converged and established the asymptotic behavior of the probability of error for the Bayes decision rule, and the majority decision rule, as a function of the number of observations.

Several closely related approaches to the problem of bounding the probability of error rely on some measure of "distance" between the distributions of the pattern classes.* The underlying decision procedure uses the likelihood ratio. The goal (which in general has not yet been reached) is to find a monotonic relation between the distance between the distributions of the pattern

*The term distance appears in quotes here because the measures do not always obey the triangle inequality required of a metric [K-1, K-2, K-5].

classe

diver-

with

Kailat

[K-2],

of one

exists

butions

bility

diverge

effecti

distribu

Th

in the t

ρ

Sev

obtained

are more

classes and the probability of error. One measure, the divergence, is defined as follows in the two-class case with the class distributions $f_1(x)$ and $f_2(x)$, respectively.

$$J = E_1\left[\ln \frac{f_1(x)}{f_2(x)}\right] - E_2\left[\ln \frac{f_1(x)}{f_2(x)}\right]$$

where

$$E_i\left[\ln \frac{f_1(x)}{f_2(x)}\right] = \int \left[\ln \frac{f_1(x)}{f_2(x)}\right] f_i(x)\, dx, \quad i = 1,2.$$

Kailath, in his well-documented paper on these techniques [K-2], points out that if one can choose the distributions of one's signal sets to increase the divergence, then there exists a set of prior probabilities under which the distributions with the larger divergence provide a lower probability of error than would distributions with a smaller divergence. However, this falls short of providing an effective computational technique for selecting signal distributions even if that option were available.

The Bhattacharyya coefficient, denoted ρ, is defined in the two-class case as

$$\rho = \int \sqrt{f_1(x) f_2(x)}\, dx.$$

Several investigators [K-2, L-2, L-3, L-4] have obtained results using the Bhattacharyya coefficient which are more encouraging than the results stemming from the

diverge

bounds

the Bha

vergenc

underly

hood de

diverge

of obse

spite o

usual s

formida

Th

method f

the conv

R-5]. R

concerni

equivoca

found tha

tial func

Raviv [H-

error. T

the proba

for any n

nique of

the asymp

thesis sh

retic poi

divergence approach. Kailath obtains both upper and lower bounds on the probability of error which are functions of the Bhattacharyya coefficient. This measure, like the divergence, indicates the potential effectiveness of the underlying probability distributions for maximum likelihood decision making. If one tries to apply either the divergence or the Bhattacharyya coefficient to a sequence of observations in a multihypothesis situation, then in spite of the simple form of the bounds, one still has the usual situation that computing the error bounds can be more formidable than computing the decision function.

The late Alfréd Rényi proposed a highly interesting method for using information theoretic measures to show the convergence of Bayes decision processes [R-3, R-4, R-5]. Rényi saw the observations as providing information concerning the classification parameter. He studied the equivocation (average entropy) of the observations and found that the equivocation approaches zero as an exponential function of the number of observations. Hellman and Raviv [H-5] related Rényi's results to the probability of error. This approach provides an absolute upper bound on the probability of error of the Bayes decision process for any number of observations. In contrast, the technique of Chu and Chueh, mentioned earlier, describes only the asymptotic behavior of the probability of error. This thesis shows the applicability of this information theoretic point of view to dependent random variables.

observ

descri

bution

distri

distri

ing di

in proc

estimat

However

sequenc

distrib

indepen

to a st

[C-2] re

mation a

as a fun

presente

random va

C.

dependenc

variation

hypothesi

elements

adjacent.

work wher

not neces

Another view of pattern recognition considers the observations as originating from a process which is described by a mixture [T-1, T-2, Y-2] of the distributions of individual component processes. The prior distribution of the classes serves as an unknown mixing distribution. Some of the approaches to estimating mixing distributions [C-4, D-1, R-7, Y-1, Y-3] have resulted in procedures which require either a steadily growing estimator space or an infinite set of finite distributions. However, Robbins [R-6] has shown how to calculate a sequence of strongly consistent estimators for the mixing distribution when the observations are statistically independent. His procedure can be shown to be equivalent to a stochastic approximation algorithm. Chien and Fu [C-2] related Bayesian learning to stochastic approximation and gave bounds on the variance of such estimators as a function of the number of observations. The work presented here applies these ideas in concert to dependent random variables.

C. K. Chow [C-6, C-7, C-8] has investigated interdependence between pattern elements (features) that is a variation on the idea of the Markov property. Chow hypothesized that the features might be ordered such that elements having interdependence would not necessarily be adjacent. This suggestion grew from character recognition work where the order in which features were measured did not necessarily reflect the history of their generation.

He co
descr
estab
resul
matin
depen
are c
a maxi
matior
bution

thesis
(which
of the
with un

M
indepen
samples
niques.
processi
the sampl
used eith
distribut
late weig
[C-11, C-
S-4]. Va
compared

He conceived a tree structure imposed on the features to describe the dependencies, and devised techniques for establishing the most effective tree. One interesting result showed that if one limits the form for approximating a probability distribution to products of low order dependent distributions; and if the low order distributions are chosen to be those for which the components exhibited a maximum mutual information; then the resulting approximation has minimum mutual information with the distribution being approximated [C-8].

There is no attempt to extend Chow's work in this thesis, since Chow worked in a supervised learning mode (which uses classified training data to evaluate parameters of the decision function), and this thesis is concerned with unsupervised techniques.

Many pattern recognition researchers have available independent, identically distributed (i.i.d.) training samples with which they can use supervised learning techniques. This avoids the high cost of optimum unsupervised processing, and no underlying probability distribution on the sample space need be specified. The training data are used either to construct approximations to the probability distributions [A-2, B-1, K-4, M-6, P-4, Y-4] or to calculate weighting parameters for discriminant functions [C-11, C-12, C-13, C-14, D-5, F-5, I-1, N-2, S-1, S-2, S-4]. Various supervised learning techniques are often compared by tuning them on the same set of training data

and th
test d
the de
would
some cr
were un
techniq
of symb
unsuperv
zation p
dependen
in the w
removed.
be appli

1.4 Cont

Thi
problem w
unsupervi
crete, dep
based on
one to si
learn the
though th
symbols f
data-gath
patterns.

and then evaluating their performance on another set of test data. Other researchers [K-4, P-4] have proved that the decision functions produced by a particular scheme would agree with the decision function which would optimize some criterion function if the number of training samples were unlimited. This thesis has not pursued supervised techniques primarily because the motivating problem, that of symbol synchronization, seemed more plausible in an unsupervised operation. Further, the symbol synchronization problem very definitely produces statistically dependent data, and the theoretically pleasing results in the works cited collapse when the i.i.d. assumption is removed. Many of the strictly empirical techniques could be applied equally well to i.i.d. or non-i.i.d. data.

1.4 Contributions of the Thesis

This thesis identifies the symbol synchronization problem with unknown symbol code length as a problem in unsupervised, multicategory pattern recognition with discrete, dependent random variables. A sequence of decisions based on a Bayes strategy is shown to converge and allow one to simultaneously determine the synchronization and learn the parameters in the source distribution. Even though the process at the source which selects successive symbols for encoding selects symbols independently, the data-gathering model at the receiver produces dependent patterns. It is pointed out that if the symbol selection

proce

and s

proce

vides

studi

cedur

variab

decisi

asympt

optimu

continu

informa

random

bility

the numb

dependen

error bo

probabil

coefficie

presents

error bou

majority

stochasti

algorithm

process at the source were Markov, then the synchronization and source parameters could still be learned by the decision process at the receiver. The binary symmetric channel provides a specific example of parameter learning which is studied. The convergence properties of such decision procedures are thoroughly examined for dependent random variables.

A suboptimum decision procedure, the majority decision procedure, is used to derive expressions for the asymptotic behavior of the probability of error for the optimum procedure; such expressions are derived for both continuous and discrete dependent random variables. An information theoretic argument, applied to the dependent random variables, provides an upper bound on the probability of error of the optimum procedure as a function of the number of observations. This thesis also extends to dependent random variables some techniques for determining error bounds which use measures of the distance between probability distributions, specifically, the Bhattacharyya coefficient, and the Kolmogorov variational distance, and presents the role of these quantities in the more elaborate error bound techniques, those using equivocation and the majority decision function. Another procedure is based on stochastic approximation of a mixing distribution, and the algorithm's convergence rate is discussed. Finally,

compu
the f
analy

1.5

in th
estab
conce
for t
of ob
which
the th
vised

computer simulations of selected techniques demonstrate the feasibility of the type of processing which has been analyzed.

1.5 Organization of the Thesis

Chapter II places the symbol synchronization problem in the framework of a pattern recognition problem and establishes convergent decision processes. Chapter III concentrates on various methods of computing error bounds for the Bayes decision process as functions of the number of observations. Chapter IV presents several examples which illustrate the theory, while Chapter V summarizes the thesis and suggests alternate approaches to unsupervised decision making.

proble

the de

will b

depend

depende

in the

eters a

it is s

arise f

specifi

the fram

describe

2.1 The

The

one of s

mation i

not rest

CHAPTER II

THE SYMBOL SYNCHRONIZATION PROBLEM

Detailed explanations of the symbol synchronization problem, the decision procedure, and the convergence of the decision procedure are presented in this chapter. It will be shown that the processing technique must deal with dependent random variables. The exact effect of this dependence on the decision process is described. While in the basic problem description the only unknown parameters are the source active and synchronization instant, it is shown that other unknown parameters, such as could arise from either a noisy channel or a less complete specification of the source, can be accommodated within the framework of the class of decision procedures described.

2.1 The Data Generation Model

The information to be received can be generated by one of several sources, and each source encodes the information in binary form. The assumption of binary data does not restrict the generality of the results obtained. Many

data

nal

and

data

a la

trans

same

sourc

binar

alpha

to an

code

more d

length

source

source,

for the

time uni

mission

its alph

and tran

necessar

locks on

source f

data transmission systems--such as remote computer terminal facilities, equipment monitors in earth satellites, and various character recognition schemes--use binary data, so the binary data model is directly applicable to a large class of existing systems.

Let L denote the number of sources. All sources transmit at the same bit rate, and all sources use the same signal to represent a binary digit. A particular source uses the same symbol code length, or number of binary digits per coded symbol, for all symbols in its alphabet. Throughout this thesis, the term symbol refers to an element of a source alphabet, and the term symbol code refers to the encoded version of a symbol. Two or more different sources might use the same symbol code length; however, it is also possible that different sources use different symbol code lengths.

By letting the letter ℓ stand for the index of the source, $\ell \in \{1,2,\ldots,L\}$, one can then have m_ℓ stand for the symbol code length used by source ℓ. Every m_ℓ time units (the time unit is the reciprocal of the transmission bit rate) source ℓ randomly selects a symbol from its alphabet, encodes the symbol using m_ℓ binary digits, and transmits the m_ℓ binary digits. At some time, not necessarily as the source begins transmitting, a receiver locks on to one of the sources and receives from that source for the rest of the time. The receiver does not

know

face

is be

lengt

each

the r

that

each

ceive

Metho

separa

binary

struct

digits

stream

$X_1, X_2, \ldots$

where n

length,

first k

stands f

tern, X_k

a.

know which source is transmitting. Two immediate tasks face the receiver. The first is to determine which source is being observed, which in turn specifies the symbol code length. It is assumed that the symbol code length of each source is known to the receiver. As a second task the receiver must determine the "synchronization instants," that is, the binary digit which is the first digit of each symbol code. Note that the first binary digit received is not necessarily the first digit of a symbol code. Methods for solving these two tasks will be developed.

The receiver observes patterns which it obtains by separating the input stream into successive sets of n binary digits. The input stream then could be reconstructed as a concatenated version of the patterns; no digits are lost in extracting the patterns from the input stream. The sequence of patterns is represented by $X_1, X_2, \ldots$ and each X_i $(i = 1,2,\ldots)$ has n binary components where n is at least as large as the largest symbol code length, i.e., $1 \leq m_\ell \leq n$. The notation X^k denotes the first k patterns; $X^k = (X_1, X_2, \ldots, X_k)$. The notation T_k stands for the synchronization instant in the k-th pattern, X_k, and is defined as follows:

a. If exactly one symbol code has its beginning binary digit in X_k, then T_k is the position of that binary digit in X_k.

b. If more than one symbol code has its beginning binary digit in X_k (which can happen whem $m_\ell < n$), then T_k is the position in X_k of the beginning digit of the first symbol code which begins in X_k.

This definition of T_k comes from the following idea: if there are exactly i binary digits (i = 0,1,...,m_ℓ-1) in X_k which belong to a symbol code which began in X_{k-1}, then T_k = i+1, and T_k tells where the first new symbol code (not left over from a symbol started in X_{k-1}) starts in X_k.

Figure 2.1 shows examples of sequences of values for T_k. Values of m_ℓ = 2 and n = 3 are assumed, and the two possible values for T_1 produce the sequences of values for $T_2,T_3,...$ as shown. Effectively then, Figure 2.1 illustrates

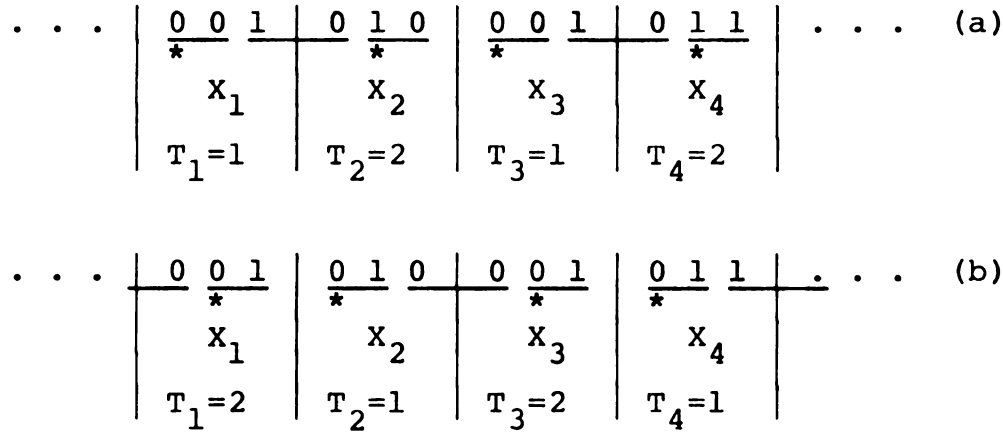


Figure 2.1 Genesis of Pattern Dependence

two ways in which the same sequence of patterns, X_k, could arise from a source which uses two binary digits for a symbol code. Pairs of underlined digits represent symbol

codes. The values of T_k are shown. The "*" under a digit shows that the digit is the one which determines the value of T_k.

The value of T_k cannot exceed the symbol code length of the source. Since n might exceed m_ℓ, T_k is not necessarily equal to T_{k+1}. However, given m_ℓ, n, and T_{k-1}, the number of digits which carry over to X_k from a symbol code started in X_{k-1} is m_ℓ minus the remainder of $\left(n - (T_{k-1} - 1)\right)/m_\ell$ so that

$$T_k = \left\{m_\ell - \left[n - T_{k-1} + 1\right]_{\text{mod } m_\ell}\right\} + 1 \qquad 2.1.1$$

The distinguishing feature of this chapter is the assumption that once the receiver has selected a stream to observe, the receiver stays with that one stream. Sections 2.2 through 2.4 develop an optimum receiver and discuss its convergence for the case in which the symbol code length and synchronization instants are the only unknowns; the probability model for the sources will be completely known and the channel between the source and receiver will be noiseless. Sections 2.5 and 2.6 allow unknown parameters in the source distributions and include a noisy channel.

2.2 Estimating the Symbol Code Length and Synchronization Instant

Each source performs a random experiment to select a symbol to be encoded. The probability of any source

selecting a given symbol is assumed known, fixed and independent of previously selected symbols. The stream of binary digits (binits) so produced represents observations on a discrete-parameter, discrete-time stochastic process. The values of the sample conditional probability mass function, for a noiseless channel

$$P(X|\ell,T_k);\ \ell = 1,2,\ldots,L;\ T_k = 1,\ldots,\ell$$

is known for all 2^n mass points and fixed. In the expression above, X--which has represented a random variable--is used to represent the value of the random variable as well. This practice will be continued, and the context will indicate whether the random variable or the value of the random variable is intended.

A symbol code might have its initial binits in X_{k-1} and its final binits at the start of X_k, which causes T_k to be greater than one. Yet no single symbol code could overlap more than two patterns, because the length of a pattern is as great as the longest code length. That being so, the value of X_k could depend on the value of X_{k-1} but not on X_j for $j < (k-1)$; recall that successive symbols are independent. Furthermore, since stationary probability models for all sources are known, the conditional probability of X_k given the value of X_{k-1} can be specified. In short, the sequence of random variables

x_1

tr

P(

and

pos

the

dend

func

k pa

$P(T_k$

quant

expre

$P(T_k$

$X_1, X_2, \ldots$ is a first-order Markov chain with stationary transition probabilities. Thus, $P(X_k|\ell,T_k,X^{k-1}) = P(X_k|\ell,T_k,X_{k-1})$.

The procedure for estimating the active source, ℓ, and the synchronization instant, T_k, is based on forming posterior estimates of the probability mass function for the active source given the first k patterns observed--denoted $P(\ell|X^k)$; $\ell = 1,\ldots,L$--and of the probability mass function for the synchronization instant given the first k patterns observed and the active source--denoted $P(T_k|\ell,X^k)$; $\ell = 1,\ldots,L$; $T_k = 1,\ldots,\ell$. Expanding these quantities by the Bayes rule gives the following recursive expressions.

$$P(\ell|X^k) = \frac{P(X_k|\ell,X^{k-1})P(\ell|X^{k-1})}{\sum_{\ell=1}^{L} P(X_k|\ell,X^{k-1})P(\ell|X^{k-1})} \qquad 2.2.1$$

$$= \frac{P(X_k|\ell,X^{k-1})P(\ell|X^{k-1})}{P(X_k|X^{k-1})} \qquad \ell = 1,2,\ldots,L$$

$$P(T_k|\ell,X^k) = \frac{P(X_k|T_k,\ell,X^{k-1})P(T_k|\ell,X^{k-1})}{\sum_{T_k=1}^{m_\ell} P(X_k|T_k,\ell,X^{k-1})P(T_k|\ell,X^{k-1})} \qquad 2.2.2$$

$$= \frac{P(X_k|T_k,\ell,X^{k-1})P(T_k|\ell,X^{k-1})}{P(X_k|\ell,X^{k-1})} \qquad \ell = 1,2,\ldots,L \quad T_k = 1,\ldots,m_\ell$$

The denominator of 2.2.2 shows how to compute the first factor in the numerator of 2.2.1. All that is left is to specify initial values for

$$P_o(\ell) = P(\ell|X^o) \text{ and}$$

$$P_o(T_k|\ell) = P(T_k|\ell,X^o)$$

because $P(X|T_k,\ell)$ is assumed fixed and known.

With the procedure for computing the posterior probabilities specified in 2.2.1 and 2.2.2 one can consider ways of estimating the value of ℓ and T_k. One estimator, called the Bayes estimator, uses the mean of the posterior distribution. The Bayes estimators of ℓ based on X^k is denoted by $\hat{\ell}_{kB}$ (the estimator for the source index after k observations) and that for T, by $\hat{T}_{kB}$ (the estimator for the synchronization instant after k observations). These estimators are defined by:

$$\hat{\ell}_{kB} = \sum_{\ell=1}^{L} \ell P(\ell|X^k)$$

and

$$\hat{T}_{kB} = \sum_{\ell} \sum_{T_k=1}^{m_\ell} T_k P(T_k|\ell,X^k) P(\ell,X^k)$$

However, these estimators tend to give fractional values for quantities defined as integers, so some rounding algorithm would have to be specified.

Another, perhaps more intuitively satisfying, estimator is the maximum likelihood estimator. The maximum likelihood estimators of ℓ and T_k based on X^k are denoted by $\hat{\ell}_{kM}$ and $\hat{T}_{kM}$, and are defined by:

$$\hat{\ell}_{kM} = \ell_o \text{ s.t. } P(\hat{\ell}_o | X^k) = \max_{\ell} P(\ell | X^k)$$

and

$$\hat{T}_{kM} = T_{ok} \text{ s.t. } P(\hat{T}_{ok} | \hat{\ell}_{kM}, X^k) = \max_{T_k} P(T_k | \hat{\ell}_{kM}, X^k).$$

2.3 Convergence of the Posterior Probability Mass Functions

The receiver is connected to a single source, ℓ_o, so all symbol codes have the same symbol code length, m_{ℓ_o}. The receiver stays connected to the source, missing none of the binary digits produced after the connection is made. Consequently, the sequence of true synchronization instants, $\{T_{o_k}\}_{k=1}^{\infty}$ will satisfy 2.1.1 with T_{o_k} substituted for T_k. It is essential to show that the posterior probability mass functions computed according to 2.2.1 and 2.2.2 converge in the sense that as $k \to \infty$ they have all of their mass at ℓ_o and T_{o_k}, respectively. In order for that to happen, a theorem of Spragins requires that the following conditions be met:

(1) the posterior probabilities must be computed by the Bayes rule,

(2) the true value of ℓ_o and T_{o_1} must have non-zero prior probabilities, and

(3) there must exist sequences of functions of the observations, $\{f_k(X^k)\}$ and $\{g_k(X^k)\}$, such that $f_k(X^k) \xrightarrow{k} \ell_o$ w.p.1. and $g_k(X^k) \xrightarrow{k} T_{o_k}$ w.p.1. where w.p.1 means "with probability one."

Equations 2.2.1 and 2.2.2 show that condition 1 is met, while proper choice of prior probabilities assures that condition 2 is met. The following theorem shows the existence of the strongly consistent estimators required in condition 3.

Theorem 2.3.1

If (a) the sequence of patterns represent observations on a regular Markov chain;

(b) only one synchronization, denoted by ℓ_o, T_{o_k} exists during the transmission of all X_k, $k = 1,2,\ldots$;

(c) members of the family $\{P(X_k|\ell,T_k);\ \ell = 1,\ldots,L;\ T_k = 1,\ldots,m_\ell\}$ are distinct;

(d) $P(X_k|\ell,T_k)$ is the same for all $k = 1,2,\ldots$

then there exist sequences of minimum distance estimators $\hat{\ell}_m$ and $\hat{T}_m$ for ℓ and T_k which converge to ℓ_o and T_{o_k} with probability one.*

*The estimators $\hat{\ell}_m$ and $\hat{T}_m$ imply a sequence of estimators $\hat{\ell}_m, \hat{T}_{1m}, \hat{T}_{2m}, \ldots, \hat{T}_{mm}, \ldots$, for which $\hat{T}_{jm}$ is not

Proof: Define the empirical discrete probability mass function for X, $\overline{P}_m(X)$, as follows:

$$\overline{P}_m(X) = \frac{1}{m} \sum_{k=1}^{m} I_X(X_k) \quad \text{a.e. where } I_X(X_k) = 1 \text{ if } X_k = X$$

$$I_X(X_k) = 0 \text{ if } X_k \neq X.$$

This is the proportion of the first m observations which equal X.

Define the estimators $\hat{\ell}_m$ and $\hat{T}_m$ by

$$\inf_{\ell, T_k} \sup_X |\overline{P}_m(X) - P(X|\ell, T_k)| = \sup_X |\overline{P}_m(X) - P(X|\hat{\ell}_m, \hat{T}_m)| \quad \text{w.p.1}$$

The infimum on the left allows us to write

$$\sup_X |\overline{P}_m(X) - P(X|\hat{\ell}_m, \hat{T}_m)| \leq \sup_X |\overline{P}_m(X) - P(X|\ell_o, T_{o_m})| \xrightarrow{m} 0 \text{ w.p.1.} \qquad 2.3.1$$

Convergence to zero with probability one is by the Glivenko-Cantelli theorem (using Signori's extension to regular Markov chains). Now the triangle inequality is used to write

---

necessarily the estimator $\hat{T}_j$ based on the j-th pattern, for $j \neq m$. Correct estimation of the synchronization means that the value of $\hat{\ell}_m$ is ℓ_o and $\hat{T}_{km}$ gives the value T_{o_k}.

$$\sup_X |P(X|\hat{\ell}_m,\hat{T}_m) - P(X|\ell_o,T_{o_m})| = \sup_X |P(X|\hat{\ell}_m,\hat{T}_m)$$

$$- \overline{P}_m(X) + \overline{P}_m(X) - P(X|\ell_o,T_{o_m})| \leq \sup_X |P(X|\hat{\ell}_m,\hat{T}_m)$$

$$- \overline{P}_m(X)| + \sup_X |\overline{P}_m(X) - P(X|\ell_o,T_{o_m})| \qquad 2.3.2$$

Equation 2.3.1 says that both quantities on the right in 2.3.2 go to zero with probability one so that consequently $P(X|\hat{\ell}_m,\hat{T}_m) \xrightarrow{m} P(X|\ell_o,T_{o_m})$ w.p.1. Hypothesis (c) then provides that

$$\hat{\ell}_m \to \ell_o \text{ w.p.1} \qquad \text{and} \qquad \hat{T}_m \to T_{o_m} \text{ w.p.1.}$$

Q.E.D.

This result in turn implies that, because of Spragins' theorem, either of the decision procedures of section 2.2 will give correct estimates of ℓ_o and T_{o_k} if the process continues long enough. With the estimators discussed so far, the rate of convergence and probability of error are difficult to specify. Section 2.4 provides an algorithm which allows statements about the rate of convergence.

2.4 A Stochastic Approximation Approach

Synchronization is completely specified by the combination of source index and synchronization instant, (ℓ, T_k).

The probability of a given pattern can be represented by the mixture

$$P(X) = \sum_{\ell=1}^{L} \sum_{T_k=1}^{m_\ell} P(X|\ell,T_k)P(\ell,T_k) \qquad 2.4.1$$

For a convenient change of notation, observe that since $1 \leqslant m_\ell \leqslant n$ and $1 \leqslant T_k \leqslant m_\ell$ there are $m = \sum_{\ell=1}^{L} m_\ell$ or at most, nL distinct pairs (ℓ,T_k) which can specify the synchronization. The notational change comes through letting an unknown parameter $\lambda \varepsilon \{1,2,\ldots,m\}$ index the possible source-synchronization pairs expressed relative to the first pattern. A sequence of functions $\gamma_k(\ell,T_k)$ is defined such that $\gamma_k(\ell,T_k)$ is a 1-1 map between the values of λ and the values of the ordered pair (ℓ,T_k). Further, $\gamma_k(\ell,T_k)$ is defined such that given the value of the pair (ℓ,T_1), the sequence of values of $(\ell,T_2),(\ell,T_3),\ldots$ generated according to 2.1.1--with ℓ held constant--maps under $\gamma_k(\ell,T_k)$ to the same value λ for $k = 1,2,\ldots$.

The parameter λ has an unknown probability vector $G = \{g_1,\ldots,g_m\}$, $g_i \geqslant 0$, $\sum_{i=1}^{m} g_i = 1$ such that $P(\lambda=i) = g_i$. Assuming that a single source with one synchronization, (ℓ_o,T_{o_k}), produces all of the observation vectors implies that one value of λ is in effect for all observations. So the system being observed has a probability vector $G = \{g_1,\ldots,g_m\}$ such that for $i = i_o$, $g_{i_o} = P(\lambda = i_o) = 1$ and for all other i, $g_i = 0$. The entry of G which has

the value l is unknown. Once i_o is known, $g_{i_o} = 1$ and all other g_i, $i \neq i_o$, must be zero.

The parameters and notation above leads to rewriting 2.4.1 as

$$P(X_k) = \sum_{a=1}^{L} \sum_{b=1}^{m_\ell} P(X_k | \ell = a, T_k = b) P(\ell = a, T_k = b)$$

$$= \sum_{i=1}^{m} P(X_k | \gamma_k(a,b) = i) P(\gamma_k(a,b) = i)$$

$$= \sum_{i=1}^{m} P_i(X_k) g_i \qquad 2.4.2$$

where $\ell = a$, $T_k = b$ maps under γ_k to $\lambda = i$,

$P_i(X_k) = P(X_k | \gamma_k = i)$, and

$P(\gamma_k(a,b) = i) = g_i$.

Estimating the synchronization now implies finding i_o, the true value of λ, for which $P(\lambda = i_o) = 1$. The approach will be to estimate the probability vector G. A decision on the value of i_o can then be made from the estimates of G.

If one can obtain a sequence of estimates $g_{i,n}$ for g_i such that $g_{i,n}$ converges to g_i, then one can use a maximum likelihood decision rule for deciding on the synchronization. The following theorem shows the

existence of a set of strongly consistent estimators for g_i (i = 1,...,m) which can be computed by a stochastic approximation algorithm (cf. [A-4]).

Theorem 2.4.1

If (a) $\{P_i(X)\}_{i=1}^m$ is an identifiable [T-1, T-2, Y-3] family of probability mass functions

(b) $G = \{g_1,...,g_m\}$, $g_i \geq 0$, $\Sigma_{i=1}^m g_i = 1$ is a finite mixing distribution, and

(c) X_k (k = 1,2,...) is a Markov chain with the distribution $P_G(X_k) = \Sigma_{i=1}^m g_i P_i(X_k)$

then there exists a sequence of estimators for the mixing distribution, G, which converges to G with probability one.

Proof: The sequence of estimators developed in the proof will have the form of a stochastic approximation algorithm. The proof follows that of H. Robbins [R-6] for a related theorem. A member of the family $\{P_i(X)\}$ will be denoted by the vector $\underline{P}_i(X)$, i = 1,...,m which has 2^n elements, one for each of the 2^n possible values of the observations X; recall that X is an n element binary vector. Yakowitz and Spragins [Y-3] show that identifiability assures that the vectors $\underline{P}_i(X)$, i = 1,...,m are linearly independent and span the space R^m (and form a basis for R^m). Let H_j denote the m-1 dimensional

subspace spanned by $\underline{P}_1(X),\ldots,\underline{P}_{j-1}(X),\underline{P}_{j+1}(X)$ $\ldots,\underline{P}_m(X)$. Then

$\underline{P}_j(X) = \underline{P}_j'(X) + \underline{P}_j''(X)$ where

$\underline{P}_j'(X) \in H_j$, $\underline{P}_j''(X) \perp H_j$ and $\underline{P}_j''(X) \neq 0$. 2.4.3

Define

$$\phi_j(X) = P_j''(X) \Big/ \sum_X [P_j''(X)]^2$$

so that

$$\sum_X \phi_j(X)P_k(X) = 1 \quad \text{if } j = k$$
$$= 0 \quad \text{if } j \neq k.$$

The elements of the set $\phi_j(X)$, $j = 1,\ldots,m$ form a set of orthonormal components of $\underline{P}_j(X)$.

Now define

$$\bar{g}_{i,n} = \frac{1}{n} \sum_{k=1}^{n} \phi_i(X_k) \quad \text{and} \quad g_{i,n} = [\bar{g}_{i,n}]^+ \Big/ \sum_{j=1}^{n} [\bar{g}_{j,n}]^+$$

where $[a]^+ = \max(a,0)$.

Hypothesis (c) and the above result imply that for all k

$$E_G\phi_i(X_k) = \sum_{X_k} \phi_i(X_k) \sum_{j=1}^{m} g_j P_j(X_k) = \sum_{j=1}^{m} g_j \sum_{X_k} \phi_i(X_k)P_j(X_k) = g_i$$

2.4.4

gonal

Gram-Sc

$P_{j-1}(X)$

an orth

zation

gonal t

H_j; thi

estima

to con

the pr

the es

for le

Signori [S-3] formulated a theorem, based on a proof of Raviv [R-2], extending the law of large numbers to Markov chains. Signori's theorem provides that when $\phi_i(X)$ is a Baire function integrable with respect to a Lebesque measure on X then 2.4.4 implies that

$\bar{g}_{i,n} \xrightarrow{n} g_i$ with probability one, hence

$g_{i,n} \xrightarrow{n} g_i$ with probability one.

Q.E.D.

In 2.4.3, the vector $\underline{P}_j''(X)$ was defined to be orthogonal to H_j; $\underline{P}_j''(X)$ can be obtained by applying the Gram-Schmidt orthogonalization procedure to $\underline{P}_1(X),\ldots,$ $\underline{P}_{j-1}(X),\underline{P}_{j+1}(X),\ldots,\underline{P}_m(X)$, which span H_j. The result is an orthogonal basis for H_j. Then the final orthogonalization step finds the portion of $\underline{P}_j(X)$ which is orthogonal to the orthogonal basis of H_j, hence orthogonal to H_j; this gives the vector $\underline{P}_j''(X)$.

While the estimators developed in Theorem 2.3.1 estimated ℓ_o and T_{o_k} and required only one source active to converge, the estimator of Theorem 2.4.1 estimates the probability distribution of the synchronization. So the estimator of Theorem 2.4.1 suggests a decision rule for learning the probability law which governs a receiver

when the receiver's data generation model selects a different source for each observation. Additional complications would be introduced in defining ℓ and T_k. Those complications would motivate the addition of data buffering to more completely state the data generation model of such a problem. It is not clear what the practical value of such a device would be.

Rewriting the definition of $\bar{g}_{i,n}$ to give it the form of a stochastic approximation algorithm begins with

$$\bar{g}_{i,n} - \bar{g}_{i,n-1} = \frac{1}{n} \sum_{k=1}^{n} \phi_i(X_k) - \frac{1}{n-1} \sum_{k=1}^{n-1} \phi_i(X_k)$$

$$= \frac{1}{n} \phi_i(X_n) + \left(\frac{1}{n} - \frac{1}{n-1}\right) \sum_{k=1}^{n-1} \phi_i(X_k)$$

$$= \frac{1}{n}[\phi_i(X_n) - \bar{g}_{i,n-1}].$$

So the recursive expression

$$\bar{g}_{i,n} = \bar{g}_{i,n-1} + \frac{1}{n}[\phi_i(X_n) - \bar{g}_{i,n-1}]$$

has the form of a stochastic approximation algorithm. Chien and Fu [C-2] show that, according to Dvoretsky's theorem, $\bar{g}_{i,n}$ converges to g_i in mean square and with probability one. Letting B denote an upper bound on the variance of $\phi_i(X_k)$, Chien and Fu show further that the mean squared error of $\bar{g}_{i,n}$ decreases at least as $\frac{1}{k}$ for k

observations and is less than or equal to B/k. This gives a bound on the rate of convergence of estimators for the mixing distribution.

2.5 Synchronization and Learning Unknown Parameters

A processor might not be fortunate enough to know the probability model for the received patterns (given the synchronization information). A likely situation would have the receiver see each source transmitting through a channel having unknown noise parameters. This section considers such unknown parameters. The functional forms of the probability distributions for the patterns, given the synchronization, are assumed known. The parameters must be learned.

The model by which data are generated is now established. Several sources generate binary data in the manner described at the beginning of this chapter. The operation of the sources is exactly as described before. At some point in time the receiver connects to the channel for one of the sources and receives from that one source through that one channel from that time forward. Each source has its own channel, such as would be the case for two space satellites whose signal paths experience different kinds of distortion, and the channel might contain noise. Each source-channel combination might contain unknown parameters which will be denoted by Θ_ℓ, where ℓ is the source index. The notation $\{\Theta_\ell\}_{\ell=1}^{L}$ denotes the

set of all unknown parameters. As before, the first bit of the first observation vector is not necessarily the first bit of a symbol code, so that the unknown synchronization instant, T_k, must be learned; T_k is defined exactly as in the introductory paragraphs of this chapter.

The received data will be processed as n binary-digit patterns denoted by X_k, k = 1,2,.... The functional form of the parameter conditional probability of the received pattern is known, so that $P(X_k|\ell,T_k,\{\Theta_\ell\},X^{k-1})$ is the fundamental known quantity from which posterior distributions will be calculated. Recursive formulas for the posterior densities follow.

$$f(\{\Theta_\ell\}|X^k) = \frac{P(X_k|\{\Theta_\ell\},X^{k-1})f(\{\Theta_\ell\}|X^{k-1})}{P(X_k|X^{k-1})}$$

$$\ell = 1,\ldots,L;\ k = 1,2,\ldots$$

$$P(\ell|\{\Theta_\ell\},X^k) = \frac{P(X_k|\ell,\{\Theta_\ell\},X^{k-1})P(\ell|\{\Theta_\ell\},X^{k-1})}{P(X_k|\{\Theta_\ell\},X^{k-1})}$$

$$\ell = 1,\ldots,L;\ k = 1,2,\ldots$$

$$P(T_k|\ell,\{\Theta_\ell\},X^k) = \frac{P(X_k|\ell,T_k,\{\Theta_\ell\},X^{k-1})P(T_k|\ell,\{\Theta_\ell\},X^{k-1})}{P(X_k|\ell,\{\Theta_\ell\},X^{k-1})}$$

$$\ell = 1,\ldots,L;\ k = 1,2,\ldots;\ T_k = 1,\ldots,m_\ell$$

2.5.1

Denominator terms come from integrating the numerator over $\{\Theta_\ell\}$, summing the numerator over ℓ, and summing the numerator over T_k, respectively, for the three equations of 2.5.1. Lower case f represents probability density function as opposed to probability mass function. Channel noise could make possible a continuum of values for components of the patterns. However, if a preprocessor were to convert the received data into binary valued pattern features, then conditional probability mass functions for the observations would be appropriate.

If ℓ_o denotes the index of the source being observed, T_o the true synchronization instant, and Θ_o the value of the parameters for the channel, then the patterns are distributed according to the conditional density

$$P(X) = P(X|\Theta_o,\ell_o,T_o),\ k = 1,2,....$$

In this case the system must learn the parameters of the channel for the source as well as the source index and synchronization instant. Spragins' theorem can be applied again to the posterior densities of 2.5.1, with the critical matter being the existence of strongly consistent estimators for Θ_o, ℓ_o, and T_o.

Theorem 2.5.1

If (a) all observations are taken from the channel connected to the source with symbol code

length m_{ℓ_o}, synchronization T_o, and channel parameters Θ_o so that

$$P(X_k) = P(X_k|\Theta_o,\ell_o,T_o) \text{ for } k = 1,2,...$$

(b) members of the family $\{P(X_k|\{\Theta_\ell\},\ell,T_k)\}_{\ell,T_k}$ are distinct

then there exist sequences of estimators $\hat{\Theta}_m$, $\hat{\ell}_m$, and $\hat{T}_m$ which converge to Θ_o, ℓ_o, and T_o, respectively, with probability one.

Proof: The proof used for Theorem 2.3.1 applies almost identically here. The empirical distribution function and the extended Glivenko-Cantelli theorem establish $P(X_k|\hat{\Theta}_m,\hat{\ell}_m,\hat{T}_m) \xrightarrow{m} P(X_k|\Theta_o,\ell_o,T_o)$ w.p.1. Hypothesis (b) then assures that the limiting $\hat{\Theta}_m$, $\hat{\ell}_m$ and $\hat{T}_m$ are unique and equal Θ_o, ℓ_o and T_o, respectively.

Q.E.D.

Notice that under this mode of operation, information is obtained about the parameters of only one channel, the channel over which the symbols are received. In general the system provides no improvement over the prior information about the parameters of the other channels.

In classical unsupervised learning [S-5] the class identification is random from observation to observation.

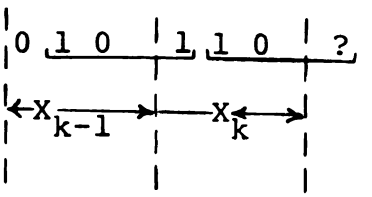
The k-th pattern there is represented by (X_k, α_k), where α_k indicates the parameters of the class being observed through the k-th pattern. The set $\{\alpha_k\}$ are usually assumed to be independent identically distributed random variables for which a set of prior probabilities, $P(\alpha_k = Z)$ must be defined. Over a long run of observations, all classes are sampled and the parameters of all classes are learned. By contrast with classical unsupervised learning, the problem considered here assumes that the class identification is fixed, though unknown.

While the formulation of problems of unsupervised learning of unknown parameters is reasonably straightforward, the computational problems in most instances are severe. Unless the functional forms are exceptionally convenient to deal with, one is forced to make discrete approximations to the continuous parameter values. This leads to the usual difficulties associated with numerical techniques, with a very difficult tradeoff between the small discretizing intervals desired for accuracy and the cost in memory and computing time that results; questions of roundoff and truncation error cannot be ignored, of course. Hellman and Cover [H-4] have worked toward a theory of the computational overhead involved in pattern recognition algorithms.

2.6 Markovian Symbol Selection

Many interesting symbol generation experiments are modeled better by assuming Markov dependence than by assuming stochastic independence among symbols. If the sources described in the single source case were modeled by first order Markov chains, then what kind of model would be appropriate for the sequence of patterns? Some insight can be gained from an example.

Assume that patterns of length n = 3 are taken from a source whose symbol code length is $m_\ell = 3$. Let $X_{k-1} = 010$, $X_k = 110$, and $T_k = 2$, as illustrated in Figure 2.2. The symbol denoted by 10? is incompletely observed in X_k and might be either 100 or 101. Since



n = 3, $X_{k-1} = 010$, $X_k = 110$, $T_k = 2$, and $m_\ell = 3$. The binary digits underlined by ⌊___⌋ are a single coded symbol. Digits through the k-th pattern are known.

Figure 2.2 Pattern Dependency with Markov Symbol Selection

the symbols are generated under first order Markovian dependence, knowledge of the previous symbol, 101, allows one to specify the probability that 10x will turn out to be 100. The distribution of X_{k+1} is conditioned on

the source, symbol code length, synchronization instant, X_k and X_{k-1}. No information prior to X_{k-1} is needed. Consequently, when the source has a symbol code length of 3 and first order Markovian dependence between symbols, the sequence of patterns of length 3, $\{X_k\}_{k=1}^{\infty}$, is also a first order Markov chain whose states correspond to the eight possible values of a pattern.

In order to generalize to all possible code lengths and pattern lengths for first order Markov sources, observe that having $T_{k+1} > 1$ caused part of X_k to be an incomplete observation of a symbol. So $m_\ell - (T_{k+1} - 1)$ digits of the incompletely observed symbol are in X_k. The symbol preceding the incompletely observed symbol has m_ℓ code digits, which, together with the $m_\ell - (T_{k+1} - 1)$ digits makes $2m_\ell - (T_{k+1} - 1) = 2m_\ell - T_k + 1$ digits which have already been observed and which have to be considered to give the conditional distribution of X_{k+1}. Since $m_\ell \leqslant n$ by assumption, $2m_\ell - T_{k+1} + 1 \leqslant 2n - T_{k+1} + 1 < 2n$ when $T_{k+1} > 1$. As a result, the patterns observed prior to X_{k-1} have no influence on the prior conditional distribution of X_{k+1}. Therefore, if the source generates symbols by using a first order Markov process, and if the source symbol code length does not exceed the observed pattern length, then the observed process can be represented by a second order Markov chain. A similar analysis can show that if the source generates symbols

by using an i-th order Markov process, and if the source code length does not exceed the observation vector length, then the observed process can be represented by a Markov process of order less than or equal to i + 1. The fact that an equivalent first order Markov process can be defined for any higher order Markov process allows one to conclude that the theorems developed for sources with independent symbols also apply to Markov sources of any order. As is well known, however, the number of states in the first order equivalent of a higher order chain increases exponentially with the order of the chain, which presents an effective restraint on the application of the technique of first order approximation. The effect of the large number of states shows up both in storage requirements and in the number of operations that must be performed.

2.7 Summary

It has been shown that the Bayes posterior distributions of the unknown parameters have a converging behavior when dependent random variables are observed and consequently the Bayes decision process is an effective procedure for deciding the symbol synchronization. In the process of proving convergence, a strongly consistent minimum distance estimator for the unknown parameters is defined, which suggests other decision procedures which are not pursued here. One alternative

decis

appro

has a

proce

zatio

of th

decis

sourc

ing,

then

tive

decision procedure which is pursued uses a stochastic approximation technique, whose estimator of the parameter has a linearly decreasing variance. The Bayes decision process can be applied simultaneously to the synchronization problem and to the problem of learning parameters of the source being observed, and the sequences of decisions on both problems will converge. Finally, if the sources use Markov processes to select symbols for encoding, rather than stochastically independent experiments, then the Bayes decision process still provides an effective solution to the symbol synchronization problem.

CHAPTER III

ERROR BOUNDS FOR DECISION PROCESSES

This chapter follows several approaches to determine the rate of convergence and probability of error of a decision procedure described in Chapter II. First, a suboptimum procedure is defined for which the asymptotic error probability can be determined. This asymptotic behavior of the error of the suboptimum procedure is used to describe the asymptotic behavior of the minimum error procedure. The second approach uses an information theoretic measure to define an upper bound on the error probability of the optimum procedure as a function of the number of observations. Subsequently, measures of the hypothesis conditional probability distributions, namely the Bhattacharyya coefficient and the Kolmogorov variational distance, give error bounds that are formally elegant but whose asymptotic properties are difficult to describe.

All of the error estimating and bounding theorems presented here require the same basic functions, the hypothesis conditional densities of the patterns, although the processing operations which are specified by each theorem

vary widely in their computational requirements. The underlying question throughout is: Is it worthwhile to compute this bound on the basis of n observations? Any worth must be measured against the computational costs, which for some of the theorems tend to offset the payoff obtained from computing the error bounds.

3.1 Majority Decision Functions and Error Probability for Dependent Random Variables

This section uses a suboptimum procedure called the majority decision procedure to obtain an upper bound on the error probability of the optimum decision procedure. Chu and Chueh [C-10] invented this approach and studied its properties for the i.i.d. case. Here, the technique will be extended to dependent random variables. An exact expression for the suboptimum error probability provides the upper bound for the optimum procedure. The asymptotic behavior of the exact expression indicates the asymptotic behavior of the optimum probability of error. Details of the application to both continuous and discrete first order dependent random variables are presented.

In an m-class decision problem, let $\Theta = \{t_1, t_2, \ldots, t_m\}$ denote the m classes, where class i occurs with probability p_i for $i = 1, \ldots, m$ and $\sum_{i=1}^{m} p_i = 1$. Let $\{X_k\}$ $k = 1, 2, \ldots$, denote a sequence of vector-valued r.v.'s, all having the same distribution. When discrete X_k are considered then $P_i(X_k)$ stands for the conditional probability

mass function for X_k when class i is active. Similarly, when continuous X_k are considered, the corresponding conditional probability density is written $f_i(X_k)$. First order dependence (1-dependence) of the X_k is assumed; i.e., $P_i(X_k|X_1,\ldots,X_{k-1}) = P_i(X_k|X_{k-1})$ for $i = 1,\ldots,m$.

3.1.1 Asymptotic Error Probability for Two Pattern Classes

In the two-class case ($m = 2$), with probabilities p_1 and p_2, let $d(X^{2n+1}) = D\left[d_1(X_1),\ldots,d_{2n+1}(X_{2n+1})\right]$ where $d_k(X_k)$ depends only on X_k so that $d_k(X_k) = t_1$ or t_2 for $k = 1,2,\ldots,2n+1$. Then $d(X^{2n+1})$ is a <u>majority decision function</u> if it follows the decision of the majority of $d_k(X_k)$. Here, $d_k(X_k)$ is a mapping from the domain of values for the k-th r.v. to the set (t_1,t_2); there can be a different mapping for each k. The function D, on the other hand, maps from the cartesian product Θ^{2n+1} to Θ. The decision regions in the observation space for X_k and the conditional probabilities of error are:

$$S_{ik} = \{X_k : d_k(X_k) = t_i\};$$

$$\alpha_{ik} = \sum_{X_k \epsilon S_{jk}} P_i(X_k) = \sum_{X_k \epsilon S_{jk}} \sum_{X_{k-1}} P_i(X_k|X_{k-1}) P_i(X_{k-1})$$

for $i \neq j$, $i,j = 1,2$, and $k = 1,2,\ldots,2n+1$.

3.1.1.1

An exact expression for the probability of error for the majority decision function will be developed with the aid of random variables which indicate the decision $d_k(X_k)$. These random variables are purposely redundant in order to provide a clear expression for the error probability and to facilitate application of the Central Limit Theorem. The approach largely follows the lead of Chu and Chueh with appropriate allowances and new definitions for handling discrete and 1-dependent random variables.

Decision indicator random variables are defined as

$$U_k = 0 \text{ and } V_k = 1 \text{ if } d_k(X_k) = t_1 \text{ ;}$$

$$U_k = 1 \text{ and } V_k = 0 \text{ if } d_k(X_k) = t_2 \text{ .} \qquad 3.1.1.2$$

In terms of these variables, the majority decision is

$$d(X^{2n+1}) = t_1 \text{ if } \sum_{k=1}^{2n+1} V_k \geq n+1$$

$$= t_2 \text{ if } \sum_{k=1}^{2n+1} U_k \geq n+1. \qquad 3.1.1.3$$

The distribution of U_k is related to the conditional error probability as follows:

$$P(U_k = 1 | \Theta = t_1) = \sum_{X_k \varepsilon S_{2k}} P_1(X_k) = \alpha_{1k}$$

and

$$P(U_k = 0|\Theta = t_1) = \sum_{X_k \epsilon S_{1k}} P_1(X_k) = 1 - \alpha_{1k}$$

so that

$$P(U_k = \xi|\Theta = t_1) = \alpha_{1k}^{\xi}(1 - \alpha_{1k})^{1-\xi}, \ \xi = 0,1$$

where α_{1k} is the probability of the decision error that $\Theta = t_2$ when in fact $\Theta = t_1$ based on X_k alone.

By a similar development,

$$P(V_k = \nu_k|\Theta = t_2) = \alpha_{2k}^{\nu}(1 - \alpha_{2k})^{1-\nu}, \nu = 0,1.$$

The exact expression for the error probability for the majority decision on 2n+1 observations is

$$P_e(d) = P(d(X^{2n+1}) = t_2, \Theta = t_1)$$

$$+ P(d(X^{2n+1}) = t_1, \Theta = t_2)$$

$$P_e(d) = p_1 P\left[\sum_{k=1}^{2n+1} U_k \geq n+1|\Theta = t_1\right]$$

$$+ p_2 P\left[\sum_{k=1}^{2n+1} V_k \geq n+1|\Theta = t_2\right] \qquad 3.1.1.4$$

$$P_e(d) = p_1 \sum^{U^*} \prod_{k=2}^{2n+1} P(U_k|\Theta = t_1, U_{k-1}) P_o(U_1|\Theta = t_1)$$

$$+ p_2 \sum^{V^*} \prod_{k=2}^{2n+1} P(V_k|\Theta = t_2, V_{k-1}) P_o(V_1|\Theta = t_2).$$

3.1.1.5

where $P_o(\cdot|\Theta)$ is the prior conditional mass function of the indicated random variable and $\Sigma^{\cdot *}$ is the sum over all sequences for which the sum of the indicated random variable exceeds n. The random variables U_k and V_k are both functions of single, 1-dependent random variables and so are, in turn, 1-dependent random variables.

Equation 3.1.1.5 can be written somewhat more compactly by defining conditional error probabilities

$$\alpha_{1k}(U_{k-1}) = P(U_k = 1|\Theta = t_1, U_{k-1})$$

and

$$\alpha_{2k}(V_{k-1}) = P(V_k = 1|\Theta = t_2, V_{k-1}).$$

Now, $P(U_k = 0|\Theta = t_1, U_{k-1}) = 1 - \alpha_{1k}(U_{k-1})$

and similarly

$$P(V_k = 0|\Theta = t_2, V_{k-1}) = 1 - \alpha_{2k}(V_{k-1}).$$

If one further defines

$$\alpha_{11}(U_o) = P_o(U_1 = 1 | \Theta = t_1) \text{ and}$$

$$\alpha_{21}(V_o) = P_o(V_1 = 1 | \Theta = t_2) \text{ then}$$

these newly defined quantities can be used in 3.1.1.5 to give

$$P_e(d) = p_1 \sum^{U^*} \prod_{k=1}^{2n+1} \alpha_{1k}^{U_k}(U_{k-1}) \left[1 - \alpha_{1k}(U_{k-1})\right]^{1-U_k}$$

$$+ p_2 \sum^{V^*} \prod_{k=1}^{2n+1} \alpha_{2k}^{V_k}(V_{k-1}) \left[1 - \alpha_{2k}(V_{k-1})\right]^{1-V_k}.$$

The form of 3.1.1.4, containing probabilities of sums of 1-dependent random variables, leads one to apply the Central Limit Theorem. In particular, consider the following factor from 3.1.1.4:

$$P\left[\sum_{k=1}^{2n+1} U_k \geq n+1 \,|\, \Theta = t_1\right].$$

In order for the Central Limit Theorem for 1-dependent random variables to apply [F-4] the following three conditions are sufficient

1. $E(U_k | \Theta = t_1) = \bar{U}_k < \infty$ must exist for $k = 1,2,\ldots$

2. $E(|U_k|^3 | \Theta = t_1) < \infty$ for $k = 1,2,\ldots$

3. $\lim_{n\to\infty} \frac{1}{n} \sum_{h=1}^{n} A_{k+h} = A$ must exist uniformly

for all k, where

$$A_k = 2\mathrm{cov}_1\{U_k, U_{k+1}\} + \mathrm{var}_1\{U_{k+1}\}$$

and where $\mathrm{cov}_i(\cdot)$ and $\mathrm{var}_i(\cdot)$ denote the covariance and variance, respectively, of the argument when class i, i = 1,2, is active. The first condition can be seen to be satisfied by considering the expectation directly,

$$E(U_k|\Theta = t_1) = 0\cdot P(U_k = 0|\Theta = t_1)$$

$$+ 1\cdot P(U_k = 1|\Theta = t_1) = \alpha_{1k}$$

$$\text{for } k = 1,2,\ldots .$$

In fact, from the expansion of the first moment, $E\{U_k|\Theta = t_1\}$, one can see that all moments and absolute moments of U_k about zero are equal to α_{1k}, so that condition 2 for the Central Limit Theorem is satisfied by the random variables $\{U_k\}$.

The expression for A_k reduces to terms containing the various conditional error probabilities already defined. The variance term is

$$\mathrm{var}_1\{U_{k+1}\} = E(|U_{k+1}|^2|\Theta = t_1) - \bar{U}_{k+1}^2$$

$$= \alpha_{1,k+1} - \alpha_{1,k+1}^2 = \alpha_{1,k+1}(1 - \alpha_{1,k+1}).$$

The covariance term is

$$\text{cov}_1\{U_k, U_{k+1}\} = E\{(U_k - \alpha_{1k})(U_{k+1} - \alpha_{1,k+1}) | \Theta = t_1\}$$

$$= E\{U_k U_{k+1} | \Theta = t_1\} - \alpha_{1k}\alpha_{1,k+1}.$$

Writing the expectation of the product,

$$E\{U_k U_{k+1} | \Theta = t_1\} = 0 \cdot 0 \cdot P(U_k = 0, U_{k+1} = 0 | t_1)$$

$$+ 0 \cdot 1 \cdot P(U_k = 0, U_{k+1} = 1 | \Theta = t_1)$$

$$+ 1 \cdot 0 \cdot P(U_k = 1, U_{k+1} = 0 | t_1)$$

$$+ 1 \cdot 1 \cdot P(U_k = 1, U_{k+1} = 1 | \Theta = t_1)$$

$$= P(U_k = 1, U_{k+1} = 1 | \Theta = t_1)$$

$$= P(U_{k+1} = 1 | \Theta = t_1, U_k = 1) \cdot$$

$$P(U_k = 1 | \Theta = t_1)$$

$$= \alpha_{1,k+1}^{(1)} \cdot \alpha_{1k} .$$

Combining the previous three equations,

$$A_k = 2\text{cov}_1\{U_k, U_{k+1}\} + \text{var}_1\{U_{k+1}\}$$

$$= 2\alpha_{1,k+1}^{(1)} \cdot \alpha_{1k} - 2\alpha_{1k}\alpha_{1,k+1} + \alpha_{1,k+1} \cdot (1 - \alpha_{1,k+1}).$$

If $P_i(X_k)$ is stationary for i = 1,2, then the $\alpha_{i,k}$ and $\alpha(j)$ are constant for all k and uniform convergence of the sequence $\left\{\frac{1}{n}\sum_{h=1}^{n} A_{k+h}\right\}_{n=1}^{\infty}$ is assured. Viewing the terms in the sequence as sample averages, one can see that the uniform convergence allows one to disregard an initial finite number of A_k and the average of the remaining A_k remains unchanged. So when the X_k are stationary, or under any other circumstance in which the sequence of sample averages on A_k converges uniformly, one can apply the Central Limit Theorem for dependent random variables [F-4] to sums of U_k to obtain

$$P\left(\sum_{k=1}^{2n+1} U_k \geq n+1 \mid \Theta = t_1\right) \rightarrow \Phi\left(n+1, \sum_{k=1}^{2n+1} \alpha_{1k}, (2n+1)A\right)$$

where A is the uniform limit described in condition 3 and Φ is defined in 3.1.1.6. A similar result applies to the sum of V_k in 3.1.1.4 so that the following theorem has been proved.

Theorem 3.1.1.1

If (a) $X_1, X_2, \ldots,$ are stationary, 1-dependent random variables under either hypothesis of a two-hypothesis decision problem,

(b) U_k and V_k are indicator functions, as defined in 3.1.1.2, for the decision based on X_k, and

(c) $\Sigma^{\phi *}$ represents the sum of the function ϕ over all sequences of length 2n+1 which give a sum $\geq$ n+1,

then the probability of error for the majority decision function after 2n+1 observations is

$$P_e(d) = p_1 \sum^{U^*} \prod_{k=2}^{2n+1} P(U_k|\Theta = t_1, U_{k-1}) P_o(U_1|\Theta = t_1)$$

$$+ p_2 \sum^{V^*} \prod_{k=2}^{2n+1} P(V_k|\Theta = t_2, V_{k-1}) P_o(V_1|\Theta = t_2)$$

$$= p_1 \sum^{U^*} \prod_{k=1}^{2n+1} \alpha_{1k}{}^{U_k(U_{k-1})} \left[1 - \alpha_{1k}(U_{k-1})\right]^{1-U_k}$$

$$+ p_2 \sum^{V^*} \prod_{k=1}^{2n+1} \alpha_{2k}{}^{V_k(V_{k-1})} \left[1 - \alpha_{2k}(V_{k-1})\right]^{1-V_k} .$$

Further,

$$\lim_{n\to\infty} P_e(d) = p_1 \Phi\left[n+1, \sum_{k=1}^{2n+1} \alpha_{1k}, (2n+1)A\right]$$

$$+ p_2 \Phi\left[n+1, \sum_{k=1}^{2n+1} \alpha_{2k}, (2n+1)B\right]$$

where

$$\Phi(x,u,\sigma^2) = \int_x^\infty (2\pi\sigma^2)^{-\frac{1}{2}} \exp[-(y-u)/2\sigma^2]dy, \qquad 3.1.1.6$$

α_{1k} and α_{2k} are defined in 3.1.1.1, A is defined in condition 3 and B is defined similarly to A, that is,

$$B = \lim_{n\to\infty} \frac{1}{n} \sum_{h=1}^{n} B_{k+h} \text{ and}$$

$$B_k = 2\text{cov}_2\{V_k, V_{k+1}\} + \text{var}_2\{V_{k+1}\}.$$

The above theorem holds whether the X_k are continuous or discrete random variables. For continuous random variables, the definitions of S_{ij}, α_{ik}, $\alpha_{1k}(U_{k-1})$ and $\alpha_{2k}(V_{k-1})$ involve appropriate integrals rather than discrete sums. However, the resulting U_k and V_k, which are used in the theorem, are discrete in either case so that the analysis above holds. Further, the theorem implies convergence of the majority decision function regardless of the local decision function used on each individual observation. If the X_k were independent, then the expression for A_k would reduce to $A_k = \alpha_{1,k+1}(1 - \alpha_{1,k+1})$ and Theorem 3 of Chu and Chueh is thereby obtained as a special case.

The asymptotic distribution of the probability of error provides a method for deciding on the number of observations required to achieve a particular level of performance. The Central Limit Theorem showed that

$$\text{Prob}\left[\Sigma U_k \geq n+1 \,|\, \Theta = t_1\right] \xrightarrow[n]{} \Phi\left[n+1, \sum_{k=1}^{2n+1} \alpha_{1k}, (2n+1)A\right]$$

3.1.1.7

and a similar statement holds for sums of the random variables V_k. If the goal is to make $P_e(d) < \xi$, then one must choose n at least large enough to force both terms of the type in 3.1.1.7 to be less than ξ, because $P_e(d)$ is the average of such terms. Defining

$$\alpha' = \sum_{k=1}^{2n+1} \alpha_{1k}$$

and

$$\beta = \frac{n+1 - \alpha'}{\sqrt{(2n+1)A}}$$

one can use tabulated values to find β such that

$$\frac{1}{2\pi} \int_{\beta}^{\infty} \exp(-x/2)dx = \xi$$

and in turn solve for n in the definition of β.

3.1.2 Bayes Majority Decision Functions

Application of the previous theorem can require substantial computing effort to evaluate the conditional error probabilities, α_{ik}, for each decision region. Restricting the decision functions at each step to a class which is formally similar to a Bayes decision produces the Bayes majority decision function, whose precise definition follows shortly. Under the Bayes majority decision function, the expressions for the mean and

variance of the limiting distribution take on a much simplified form. This simplified form contains a factor which is derived from an upper bound on the Kolmogorov variational distance between the class distributions.

In what follows, one needs to use a particular property of the probability of error of individual decisions in a two-hypothesis decision problem. Using essentially the same notation as above, for a discrete random variable X let $P_1(\cdot)$ and $P_2(\cdot)$ denote the probability mass functions for X given $\Theta = t_1$ and $\Theta = t_2$, respectively. Also let

$$S_1 = \{x : d(x) = t_1\} \quad \text{and} \quad S_2 = S_1'.$$

Then

$$\alpha_1 = \sum_{x \in S_2} P_1(x) \quad \text{and} \quad \alpha_2 = \sum_{x \in S_1} P_2(x)$$

are the conditional probabilities of error for rule d. The overall probability of error, P_e, is then

$$P_e = p_1\alpha_1 + p_2\alpha_2 \qquad \text{where} \qquad p_1 + p_2 = 1.$$

This implies that $\min(\alpha_1, \alpha_2) \leq P_e \leq \max(\alpha_1, \alpha_2)$ with equality holding only if $\alpha_1 = \alpha_2$, in which case $\alpha_1 = \alpha_2 = P_e$. Define

$$\underline{\alpha} \equiv \min(\alpha_1, \alpha_2),$$

$$\overline{\alpha} \equiv \max(\alpha_1, \alpha_2) \quad \text{and}$$

$$\eta \equiv \overline{\alpha} - \underline{\alpha} = |\alpha_1 - \alpha_2|$$

so that η denotes the length of the interval within which P_e is bounded.

Theorem 3.1.2.1

If $\sum_x |P_1(x) - P_2(x)| \geq 2\delta$ then $P_e \leq \frac{1}{2} - \frac{\delta}{4} + \frac{\eta}{2}$.*

Proof: Case 1: $p_1 \geq p_2$. By steps which are identical to those in Theorem 2 of Chu and Chueh [C-10] one gets

$$1 - P_e \geq p_1\delta + p_2\alpha_2 + p_1\alpha_1 = p_1\delta + P_e.$$

Substituting a lower bound for P_e on the right gives

$$1 - P_e \geq p_1\delta + \underline{\alpha}$$

and the P_e term can be isolated on the right by

$$1 - p_1\delta - \underline{\alpha} \geq P_e.$$

---

*The quantity $\sum_x |P_1(x) - P_2(x)|$, or $\int_x |f_1(x) - f_2(x)|dx$ when x is a continuous r.v., is the Kolmogorov variational distance.

Next, add $\bar{\alpha}$ to both sides

$$1 - p_1\delta + \bar{\alpha} - \underline{\alpha} \geq P_e + \bar{\alpha}.$$

But $P_e + \bar{\alpha} \geq 2\ P_e$ and $\eta = \bar{\alpha} - \underline{\alpha}$ so that

$$1 - p_1\delta + \eta \geq 2\ P_e$$

or

$$P_e \leq \tfrac{1}{2}(1 - p_1\delta + \eta)$$

and if $p_1 \geq p_2$ then $p_1 \geq \frac{1}{2}$ so $P_e \leq \frac{1}{2}(1 - \frac{\delta}{2} + \eta)$.

Case 2: $p_1 < p_2$. In this case $1 - P_e \geq p_2\delta + P_e$ and an analysis of the type above with p_1 replaced by p_2 again produces the conclusion of the theorem.

Q.E.D.

Still considering the two-class decision problem, a Bayes majority decision function is a majority decision function such that for every $k = 1,2,\ldots,2n+1$, the decision regions, S_{ik}, are defined as

$$S_{1k} = \{x_k : q_{1k}P_{1k}(x_k) \geq q_{2k}P_{2k}(x_k)\} \text{ and } S_{2k} = S_{1k}'$$

where $q_{1k}, q_{2k} \geq 0$ and $q_{1k} + q_{2k} = 1$. The q_{ik} need not be the true probability, p_i, that $\Theta = t_i$, nor need they be the Bayes posterior estimates of p_i.

The least favorable distribution of Θ with respect to $P_{1k}(\cdot)$ and $P_{2k}(\cdot)$ is defined to be that set of values for q_{1k} and q_{2k} which minimizes $\eta_k = \overline{\alpha}_k - \underline{\alpha}_k$, where $\overline{\alpha}_k = \max(\alpha_{1k},\alpha_{2k})$ and $\underline{\alpha}_k = \min(\alpha_{1k},\alpha_{2k})$ and α_{1k} is defined in 3.1.1.1. This distribution is least favorable in the sense that, since $\underline{\alpha}_k$ is a lower bound on $P_e\left(d_k(x_k)\right)$, and since minimizing η_k implies maximizing $\underline{\alpha}_k$ (because α_{1k} and α_{2k} are measures of complementary regions in the sample space), then minimizing η_k maximizes the lower bound which $\underline{\alpha}_k$ places on $P_e\left(d_k(x_k)\right)$. However, minimizing η_k does minimize the general upper bound of the previous theorem. For continuous random variables, the minimum value of η_k is zero, provided that $f_1(x) \not\equiv f_2(x)$.

In the case of continuous random variables, the following theorem results:

Theorem 3.1.2.2

If (a) x^{2n+1} is a sequence of 1-dependent, continuous random variables;

(b) if $d(x)$ is the Bayes majority decision function such that for $k = 1,\ldots,2n+1$, q_{1k} and q_{2k} are the least favorable distribution of Θ with respect to $f_{1k}(x_k)$ and $f_{2k}(x_k)$; and

(c) if for every k, $\int |f_{1k} - f_{2k}| \geq 2\delta > 0$

then

$$\lim_{n\to\infty} P_e(d) \leq \lim_{n\to\infty} \Phi\left(n+1,(2n+1)\varepsilon,3(2n+1)\varepsilon(1-\varepsilon)\right) \quad \text{w.p.1}$$

where

$\varepsilon = \frac{1}{2} - \frac{\delta}{4}$. Consequently, $P_e(d) \to 0$ as $\eta \to \infty$, w.p.1

Proof: The least favorable distribution for continuous random variables makes $\alpha_{1k} = \alpha_{2k}$ and $\eta = 0$, so by the above theorem $\alpha_{ik} \leqslant \varepsilon < \frac{1}{2}$, $i = 1,2$. The α_{ik} can be replaced by ε in Theorem 3.1.1.1 and the variance terms are bounded above by $3\varepsilon(1-\varepsilon)$. That is,

$$A_k = 2\alpha_{1k}(\alpha_{1,k+1}(1) - \alpha_{1,k+1}) + \alpha_{1,k+1}(1 - \alpha_{1,k+1})$$

$$\leqslant 2\varepsilon(1-\varepsilon) + \varepsilon(1-\varepsilon) = 3\varepsilon(1-\varepsilon).$$

Consequently

$$\lim_{n\to\infty} \frac{1}{n} \sum_{h=1}^{n} A_{k+h} \leqslant 3\varepsilon(1-\varepsilon).$$

A similar approach holds for B_k and B. Thus, both normal distribution functions in 3.1.1.6 are bounded above by

$$\Phi\left[n+1, (2n+1)\varepsilon, (2n+1) \cdot 3\varepsilon(1-\varepsilon)\right] = \Phi\left[\frac{n+1 - (2n+1)\varepsilon}{\sqrt{3(2n+1)\varepsilon(1-\varepsilon)}}, 0, 1\right].$$

Since $\varepsilon < \frac{1}{2}$, the first argument is positive and increases as $n^{\frac{1}{2}}$ so the theorem follows.

Q.E.D.

One encounters difficulty in attempting to apply the above proof technique to a similar theorem for discrete random variables. In particular, since $\eta \neq 0$ in the definition of least favorable distribution, it is not necessarily the case that proper choice of q_{1k} and q_{2k} can force both $\alpha_{ik} \leqslant \frac{1}{2}$ for i = 1,2. As an example, consider the following two discrete distributions on {0,1}.

$$P_1(0) = 0.9 \qquad P_2(0) = 0.8$$

$$P_1(1) = 0.1 \qquad P_2(1) = 0.2$$

The full set of possible partitionings of the sample space into decision regions, and the resulting conditional error probabilities are:

	$d^1(x_k)$	$d^2(x_k)$	$d^3(x_k)$	$d^4(x_k)$
S_1	{0,1}	{0}	{1}	{ϕ}
S_2	{ϕ}	{1}	{0}	{0,1}
α_1	0.0	0.1	0.9	1.0
α_2	1.0	0.8	0.2	0.0

For this example, no decision function exists for which both α_1 and α_2 are less than $\frac{1}{2}$. Consequently, substitution

of an upper bound for α_{ik}, i = 1,2, in 3.1.1.6 is fruitless.

However, for sample spaces which can be partitioned so that $\alpha_i < \frac{1}{2}$, i = 1,2, the above theorem extends to discrete 1-dependent distributions.

3.1.3 Asymptotic Error Probability for m Pattern Classes

In the m-hypothesis case, $\Theta = \{t_1,\ldots,t_m\}$, the maximum likelihood-Bayes decision regions for a single observation are defined as

$$T_i = \{x : p_i f_i(x) = \max_{j=1,\ldots,m} p_j f_j(x)\},\ i = 1,\ldots,m$$

and the pairwise decision regions are

$$S_{ij} = \{x : p_i f_i(x) \geq p_j f_j(x)\} \quad i,j = 1,\ldots,m.$$

In the expressions above, $f_i(x)$ is the conditional probability density of x given $\Theta = t_i$. The error probability for the Bayes decision procedure is

$$P_e(\text{Bayes}) = \sum_{i<j} \left[\int_{T_i} p_j f_j(x)dx + \int_{T_j} p_i f_i(x)dx \right].$$

Since $T_i \subseteq S_{ij}$ and $T_j \subseteq S'_{ij}$, it follows that

$$P_e(\text{Bayes}) \leq \sum_{i<j} \left[\int_{S_{ij}} p_j f_j(x)dx + \int_{S'_{ij}} p_i f_i(x)dx \right].$$

Consider the problem of deciding $\Theta = t_i$ vs. $\Theta \neq t_i$ by a Bayes majority decision function when there are m classes. That is, m such two-class partitionings can be made. Under the hypothesis $\Theta = t_i$, x has the density $f_i(x)$. Under the composite hypothesis $\Theta \neq t_i$, x has the density

$$f_{\hat{i}}(x) \equiv \frac{1}{1 - p_i} \sum_{\substack{j=1 \\ j \neq i}}^{m} p_i f_j(x).$$

Let

$$S_{ik} = \{x_k : q_{ik} f_{ik}(x_k) \geq q_{\hat{i}k} f_{\hat{i}k}(x_k)\} \text{ and } S_{\hat{i}k} = S'_{ik}$$

where $q_{ik}, q_{\hat{i}k} \geq 0$ and $q_{ik} + q_{\hat{i}k} = 1$. The decision for x_k is: Decide $\Theta = t_i$ if $x_k \varepsilon S_{ik}$ and decide $\Theta \neq t_i$ otherwise. Then the above theorem says that if for all k, $\int |f_{ik} - f_{\hat{i}k}| \geq 2\delta > 0$ then

$$\lim_{n \to \infty} P_e(d) \leq \lim_{n \to \infty} \Phi\left(n+1, (2n+1)\varepsilon, (2n+1)\varepsilon(1-\varepsilon)\right)$$

where $\varepsilon = \frac{1}{2} - \frac{\delta}{4}$. Define the <u>compound Bayes majority decision procedure</u>, d_{CB}, by making Bayes majority decisions for the m decisions $\Theta = t_i$ vs. $\Theta \neq t_i$, i = 1,...,m, and deciding $\Theta = t_j$ where j is the value of i for which the two-way Bayes majority decision was $\Theta = t_i$, provided that exactly one such j exists. Otherwise, no compound decision is made, but another observation would be taken.

Theorem 3.1.3.1

If (a) x is a sequence of 1-dependent observations in an m-class decision problem, and

(b) if for all i = 1,...,m and k = 1,2,...,

$$\int |f_{ik} - f_{\hat{i}k}| \geq 2\delta > 0$$

then

$$\lim_{n\to\infty} P_e(d_{CB}) \leq (m-1) \lim_{n\to\infty} \left[\Phi\left(n+1, (2n+1)\varepsilon, 3(2n+1)\varepsilon(1-\varepsilon)\right)\right]^2 \quad \text{a.e.}$$

where $\varepsilon = \frac{1}{2} - \frac{\delta}{4}$

Consequently $\lim_{n\to\infty} P_e(d_{CB}) = 0$ a.e.

Proof: Of the decisions which d_{CB} can make, m-1 represent errors, and each of those results from exactly two errors in the two-way Bayes majority decision procedures. Since the above theorem applies to the probability of error of the Bayes majority decision procedures, the result follows.

Corollary: Under the same hypotheses

$$\lim_{n\to\infty} P_e(\text{Bayes}) \leq (m-1) \lim_{n\to\infty} \left[\Phi\left(n+1, (2n+1)\varepsilon, 3(2n+1)\varepsilon(1-\varepsilon)\right)\right]^2 \quad \text{a.e.}$$

Proof: The Bayes decision procedure minimizes the error probability, so the error of any other decision procedure, such as d_{CB}, provides an upper bound for $P_e(\text{Bayes})$.

3.2 Bounding the Probability of Error After a Finite Number of Samples

In studying m-class hypothesis testing using dependent random variables, the results presented so far have shown the existence of a strongly consistent minimum distance estimator for the pattern class, that Bayes rule calculations of posterior distributions converge in the sense that eventually the mass of the posterior distribution lies at the point corresponding to the correct class index, and that the error probability of a suboptimum procedure--and consequently of the Bayes procedure--vanishes as the number of observations grows without bound. Now it will be shown that an upper bound exists for the probability of error of the Bayes decision process, and that the upper bound decreases exponentially as the number of observations increases. An exact expression for the upper bound is obtained as a function of the number of patterns observed. The expression is derived from information theoretic considerations, and it turns out that the interclass Bhattacharyya coefficients are factors of terms in the bound.

This presentation will use the general notation which applies to the class of m-hypothesis decision-making problems for which there is first order dependence between successive observations. The symbol synchronization problem with unknown source code length is a member of that class. For convenient reference, the specification

of the m-class decision-making problem and the notation to be used will be briefly reviewed. This is the same problem and notation which was described in detail earlier.

Let $X_1, X_2, \ldots, X_k, \ldots$ be a sequence of identically distributed discrete* random variables having first order dependence. The common distribution of X_k, $k = 1,2,\ldots,$ depends on the pattern class, indexed by $\Theta = \{t_1, t_2, \ldots, t_m\}$ and the prior probability for the pattern class j is given by $\text{Prob}(\Theta = t_j) = p_j > 0$. The parameter conditional distribution of the X_i will be written

$$\text{Prob}\{X_k | \Theta = t_j\} = P_j(X_k), \quad j = 1,\ldots,m; \quad k = 1,2,\ldots,$$

and the first order dependence provides that

$$P_j(X_k | X_1, X_2, \ldots, X_{k-1}) = P_j(X_k | X_{k-1}).$$

It is assumed that each class has a unique probability distribution; that is, $P_j(X_k | X_{k-1})$ and $P_h(X_k | X_{k-1})$ are not identical for all values of the arguments when $j \neq h$.

3.2.1 An Information Theoretic Approach

After k observations, the amount of information contained in the sequence of random variables

---

*The arguments to be set forth would apply to continuous random variables as well.

$X^k = X_1, X_2, \ldots, X_k$ about Θ is the mutual information

$$I_k = I_k(X^k, \Theta) = H(\Theta) - E\left(H(\Theta|X^k)\right),$$

where $H(\Theta)$ is Shannon's entropy:

$$H(\Theta) = \sum_{j=1}^{m} p_j \log\left(\frac{1}{p_j}\right)$$

and $p_j = \text{Prob}(\Theta=t_j)$ as described above. Similarly

$$H(\Theta|X^k) = \sum_{j=1}^{m} P\left(\Theta=t_j|X^k\right)\log\left(\frac{1}{P(\Theta=t_j|X^k)}\right).$$

Logarithms are to the base 2, and throughout this chapter, $E(\cdot)$ denotes the expectation with respect to the joint distribution of X^k.

The development which follows is based on work done in 1964 by A. Rényi [R-3]. Rényi described the behavior of the average entropy for independent random variables and used his results to show the almost sure convergence of a decision procedure which was similar to the maximum likelihood decision procedure. Hellman, Raviv and others have called the expectation of the entropy $E\left(H(\Theta|X^k)\right)$, the <u>equivocation</u>. Hellman and Raviv [H-5] showed that for the Bayes decision procedure, the probability of error, $P_B(e)$, is bounded above by one-half the equivocation, i.e.,

$$P_B(e) \leq \frac{1}{2}E\left(H(\Theta|X^k)\right).$$

They showed also that in the i.i.d. case,

$$E\left(H(\Theta|X^k)\right) \leq K(\rho^{**})^k$$

where ρ^{**} is defined as follows.

$$\rho^*_{ij} = \inf_{0\leq\alpha\leq 1} \sum_{x} P_i(x)^{\alpha} P_j(x)^{1-\alpha} \quad \text{and}$$

$$\rho^{**} = \max_{i\neq j} \rho^*_{ij}$$

When $\alpha = \frac{1}{2}$, the argument of the infimum is called the interclass Bhattacharyya coefficient, so that Hellman and Raviv's result for i.i.d. random variables defines an upper bound on the error probability of the Bayes decision procedure which is an exponentially decreasing function of the maximum interclass Bhattacharyya coefficient. In order to obtain related results for dependent random variables, two lemmas must be established.

Lemma 3.2.1.1

(Rényi) A universal constant $C > 0$ exists such that for any set $p_1,\ldots,p_m$ of positive numbers forming a probability distribution

$$H(\Theta) = \sum_{j=1}^{m} p_j \log\left(\frac{1}{p_j}\right) \leqslant C \sum_{j=2}^{m} p_j^{\alpha} \quad \text{for } 0 \leqslant \alpha \leqslant 1.$$

The logarithm is to the base 2.

Proof: This is Rényi's proof, although he used $\alpha = \frac{1}{2}$.

Both $\dfrac{x \log\left(\frac{1}{x}\right)}{x^{\alpha}}$ and $\dfrac{(1-x)\log\left(\frac{1}{1-x}\right)}{x^{\alpha}}$ are continuous

in $[0,1]$.

Define

$$C_1 = \max_{0 \leqslant x \leqslant 1} \frac{x\log\left(\frac{1}{x}\right)}{x^{\alpha}} \text{ and } C_2 = \max_{0 \leqslant x \leqslant 1} \frac{(1-x)\log\left(\frac{1}{1-x}\right)}{x^{\alpha}}$$

for $0 \leqslant \alpha \leqslant 1$

Then by breaking the entropy expression into two parts, one gets bounds on each part of

$$\sum_{j=2}^{m} p_j \log\left(\frac{1}{p_j}\right) = \sum_{j=2}^{m} \frac{p_j \log\left(\frac{1}{p_j}\right)}{p_j^{\alpha}} p_j^{\alpha} \leqslant C_1 \sum_{j=2}^{m} p_j^{\alpha}$$

and

$$p_1 \log \frac{1}{p_1} = \left[1 - \sum_{j=2}^{m} p_j\right] \log \left[\frac{1}{1 - \sum_{j=2}^{m} p_j}\right]$$

$$= \frac{\left[1 - \sum_{j=1}^{m} p_j\right] \log \left[\frac{1}{1 - \sum_{j=2}^{m} p_j}\right]}{\left[\sum_{j=2}^{m} p_j\right]^{\alpha}} \left[\sum_{j=2}^{m} p_j\right]^{\alpha}$$

$$\leq C_2 \left[\sum_{j=2}^{m} p_j\right]^{\alpha} \leq C_2 \sum_{j=2}^{m} p_j^{\alpha}$$

The lemma follows with $C = C_1 + C_2$.

Q.E.D.

Lemma 3.2.1.2

If $Y_1, Y_2, \ldots, Y_k, \ldots$ is a sequence of random variables having the Markov property, and if $E(Y_k|Y_{k-1})$ exists and is bounded for all k, say

$$E(Y_k|Y_{k-1}) \leq K \quad , \quad k = 2,3,\ldots \quad \text{w.p.1}$$

then

$$E\left(\sum_{i=1}^{k} Y_i\right) \leq E(Y_1)K^{k-1} \quad \text{where } K = E(Y_2|Y_1).$$

Proof: Expanding the expression for the expectation of the product of k random variables having the Markov property gives

$$E\left(\prod_{i=1}^{k} Y_i\right) = \sum_{Y_1} \cdots \sum_{Y_k} Y_1 \cdot Y_2 \cdots Y_k P(Y_k|Y_{k-1}) \cdots P(Y_2|Y_1) P(Y_1|Y_0)$$

$$= \sum_{Y_1} \cdots \sum_{Y_{k-1}} Y_1 \cdots Y_{k-1} \sum_{Y_k} Y_k P(Y_k|Y_{k-1}) P(Y_{k-1}|Y_{k-2}) \cdots$$

$$P(Y_2|Y_1) P(Y_1|Y_0)$$

$$= \sum_{Y_1} \cdots \sum_{Y_{k-1}} Y_1 \cdots Y_{k-1} E(Y_k|Y_{k-1}) P(Y_{k-1}|Y_{k-2}) \cdots$$

$$P(Y_2|Y_1) P(Y_1|Y_0)$$

$$\leq \sum_{Y_1} \cdots \sum_{Y_{k-1}} Y_1 \cdots Y_{k-1} K P(Y_{k-1}|Y_{k-2}) \cdots P(Y_2|Y_1) P(Y_1|Y_0)$$

This process repeats k-2 more times to yield the lemma.

Q.E.D.

These lemmas will be used in proving the following theorem.

<u>Theorem 3.2.1.1</u>

If (a) Θ is a discrete random variable taking on m different values $t_1, t_2, \ldots, t_m$ with positive prior probabilities $p_j = \text{Prob}(\Theta = t_j)$, $j = 1,2,\ldots,m$;

(b) the discrete random variables $X_1, X_2, \ldots, X_k, \ldots$ have, for each j, identical conditional, given

$\Theta = t_j$, distributions with the Markov property; and

(c) the conditional joint distributions of X_{k-1}, X_k given $\Theta = t_j$ versus $\Theta = t_h$ are different for each $j \neq h$, i.e., there is no value of X_{k-1} such that $P_h(X_k|X_{k-1}) \equiv P(X_k|X_{k-1})$,

then there exist positive constants A and $q < 1$ such that

$$0 \leqslant E\left(H(\Theta|X^k)\right) \leqslant Aq^{k-1} \quad \text{for } k = 1,2,\ldots$$

where $X^k = X_1, X_2, \ldots, X_k$.*

Proof: Letting Ω_h denote the subset of the full probability space, Ω, on which $\Theta = t_h$ $(h = 1,\ldots,m)$, the equivocation is expanded in terms of the parameter conditional equivocation as follows.**

$$E\left(H(\Theta|X^k)\right) = \sum_{h=1}^{m} p_h E\left(H(\Theta|X^k)|\Omega_h\right) \qquad 3.2.1.1$$

The Bayes posterior distribution of Θ given X^k, which is needed to evaluate the entropy factors in 3.2.1.1 is:

---

*Korsh [K-6] proved a similar theorem using a different proof.

**The notations $E(\cdot|\Omega_h)$ and $E(\cdot|\Theta = t_h)$ mean the same thing. They are alternative notations for the same conditional expectation.

$$P(\Theta = t_j | X^k) = \frac{p_j \prod_{i=1}^{k} P_j(X_i | X_{i-1})}{\sum_{h=1}^{m} p_h \prod_{i=1}^{k} P_h(X_i | X_{i-1})}$$

$$< \frac{p_j}{p_h} \prod_{i=1}^{k} \frac{P_j(X_i | X_{i-1})}{P_h(X_i | X_{i-1})} \qquad 3.2.2.2$$

where the strict inequality holds for all finite k and $h = 1,2,\ldots,m$. Here, $P_j(X_1 | X_0)$ stands for the prior conditional probability of X_1, the first pattern.

Now the entropy expression can be expanded, Lemma 3.2.1.1 applied to the expansion, and then the bound in 3.2.1.2 applied to the result:

$$H(\Theta | X^k) = \sum_{j=1}^{m} P(\Theta = t_j | X^k) \log \left(\frac{1}{P(\Theta = t_j | X^k)} \right)$$

$$\leqslant C \sum_{\substack{j=1 \\ j \neq h}}^{m} \left(P(\Theta = t_j | X^k) \right)^{\alpha}$$

$$< C \sum_{\substack{j=1 \\ j \neq h}}^{m} \left(\frac{p_j}{p_h} \right)^{\alpha} \prod_{i=1}^{k} \left(\frac{P_j(X_i | X_{i-1})}{P_h(X_i | X_{i-1})} \right)^{\alpha}$$

$$\text{for } 0 \leqslant \alpha \leqslant 1. \qquad 3.2.1.3$$

The function $P_j(X_i | X_{i-1})$ depends on the random variable X_{i-1}, so one can take the expectation of the

bound in 3.2.1.3 given Ω_h, recalling that Ω_h specifies the distribution of X^k. The expectation wanted is

$$E\left(H(\Theta|X^k)\,|\,\Omega_h\right) < C \sum_{\substack{j=1 \\ j\neq h}}^{m} \left(\frac{p_j}{p_h}\right)^{\alpha} E\left\{\prod_{i=1}^{k}\left(\frac{P_j(X_1|X_{i-1})}{P_h(X_i|X_{i-1})}\right)^{\alpha}\middle|\,\Omega_h\right\}.$$

3.2.1.4

The rest of the proof proceeds using $\alpha = \frac{1}{2}$ for the sake of a helpful cancellation which develops. Lemma 3.2.1.2 can be applied to the $E\{\ldots\}$ on the right of 3.2.1.4 as soon as the K bound is demonstrated. In particular, it will be shown that

$E\left\{\left(\frac{P_j(X_2|X_1)}{P_h(X_2|X_1)}\right)^{\frac{1}{2}}\middle|\,\Omega_h\right\} < 1$ to give a bound on the equivocation which decreases monotonically with k.

$$E\left\{\left(\frac{P_j(X_2|X_1)}{P_h(X_2|X_1)}\right)^{\frac{1}{2}}\middle|\,\Omega_h\right\} = \sum_{X_1,X_2}\left(\frac{P_j(X_2|X_1)}{P_h(X_2|X_1)}\right)^{\frac{1}{2}} P_h(X_2,X_1)$$

$$= \sum_{X_1,X_2}\left(\frac{P_j(X_2,X_1)\,|\,P_j(X_1)}{P_h(X_2,X_1)\,|\,P_h(X_1)}\right)^{\frac{1}{2}} P_h(X_2,X_1)$$

$$= \sum_{X_1,X_2}\left(\frac{P_h(X_1)}{P_j(X_1)}\right)^{\frac{1}{2}}\left(P_j(X_2,X_1)P_h(X_2,X_1)\right)^{\frac{1}{2}}$$

$$= \sum_{X_1}\left(\frac{P_h(X_1)}{P_j(X_1)}\right)^{\frac{1}{2}} \sum_{X_2}\left(P_j(X_2,X_1)P_h(X_2,X_1)\right)^{\frac{1}{2}}$$

$$< \sum_{X_1} \left(\frac{P_h(X_1)}{P_j(X_1)} \right)^{\frac{1}{2}} \left[\sum_{X_2} P_j(X_2, X_1) \sum_{X_2} P_h(X_2, X_1) \right]^{\frac{1}{2}}$$

3.2.1.5

The inequality in 3.2.1.5 is an application of the Schwarz inequality. Strict inequality is a result of the assumption of unique parameter conditional distributions, i.e., the condition for which equality would hold in the Schwarz inequality is equivalent to $P_j(X_2|X_1) \equiv P_h(X_2|X_1)$ for some j and h, $j \neq h$. The right side of 3.2.1.5 reduces to 1, so there exists a q_{jh}, $0 \leq q_{jh} < 1$, such that

$$E\left\{ \left[\frac{P_j(X_2|X_1)}{P_h(X_2|X_1)} \right]^{\frac{1}{2}} \middle| \Omega_h \right\} = q_{jh} < 1 \qquad 3.2.1.6$$

Letting $\alpha = \frac{1}{2}$ in 3.2.1.4 and applying Lemma 2 with $K = q_{jh}$ results in 3.2.1.4 becoming

$$E\left(H(\Theta|X^k)|\Omega_h\right) < C \sum_{\substack{j=1 \\ j \neq h}}^{m} \left(\frac{p_j}{p_h} \right)^{\frac{1}{2}} E\left\{ \left[\frac{P_j(X_1)}{P_h(X_1)} \right]^{\frac{1}{2}} \middle| \Omega_h \right\} q_{jh}^{k-1} \, .$$

Also, 3.2.1.7

$$E\left\{ \left[\frac{P_j(X_1)}{P_h(X_1)} \right]^{\frac{1}{2}} \middle| \Omega_h \right\} = \sum_{X_1} \left[\frac{P_j(X_1)}{P_h(X_1)} \right]^{\frac{1}{2}} P_h(X_1) = \sum_{X_1} \left[P_j(X_1) P_h(X_1) \right]^{\frac{1}{2}}$$

$$= \rho_{jh} < 1 \qquad 3.2.1.8$$

This expectation turns out to be precisely the inter-class Bhattacharyya coefficient [H-5, K-2, L-3] for the prior distributions of the first observation, X_1.

As in 3.2.1.5, the strict inequality in 3.2.1.8 is from the Schwarz inequality and the unique parameter conditional distributions. Now defining $q \triangleq \max_{1 \leq j<h} (q_{jh}) < 1$ and using this q with 3.2.1.7 and 3.2.1.8 in 3.2.1.1 gives

$$E\left(H(\Theta|X^k)\right) < \sum_{h=1}^{m} p_h \, C \sum_{\substack{j=1 \\ j\neq h}}^{m} \left(\frac{p_j}{p_h}\right)^{\frac{1}{2}} \rho_{jh} q^{k-1} \qquad 3.2.1.9$$

This proves the theorem with

$$A = C \sum_{h=1}^{m} \sum_{\substack{j=1 \\ j\neq h}}^{m} (p_j p_h)^{\frac{1}{2}} \rho_{jh} < C(m-1) \qquad 3.2.1.10$$

Q.E.D.

In Rényi's derivation for the case of independent random variables q turns out to be the maximum interclass Bhattacharyya coefficient.

Corollary: Under the hypotheses of the theorem, the probability of error for the Bayes decision procedure is bounded by

$$P(e) \leq \frac{1}{2} A q^{k-1} \quad , \quad k = 1,2,\ldots . \qquad 3.2.1.11$$

where A is defined in 3.2.1.10 and q is the maximum of the q_{jh} defined in 3.2.1.6.

Proof: Hellman and Raviv [H-5] showed that the error probability for the Bayes decision procedure is bounded by $\frac{1}{2}$ the equivocation which this theorem shows is bounded in turn by Aq^{k-1}.

Q.E.D.

In the argument between 3.2.1.4 and 3.2.1.5, a value of $\alpha = \frac{1}{2}$ was chosen in order to argue that $q_{jh} < 1$ in 3.2.1.6. With that accomplished, one can establish a tighter bound on the probability of error by defining, as an alternative to q_{jh} in 3.2.1.6,

$$q'_{jh} = \inf_{0<\alpha\leq 1} E\left\{\left[\frac{P_j(X_2|X_1)}{P_h(X_2|X_1)}\right]^{\alpha} \middle| \Omega_h\right\}$$

Since the definition of q_{jh} exhibits a value of α, $\alpha = \frac{1}{2}$, for which the expectation is less than 1, then

$$q'_{jh} \leq q_{jh} < 1$$

Letting $\bar{\alpha}$ denote the value of α for which the infimum is obtained, and following the subsequent steps of the theorem gives an alternative to 3.2.1.9.

$$E\left(H(\Theta|X^k)\right) < \sum_{h=1}^{m} p_h C \sum_{\substack{j=1 \\ j\neq h}}^{m} \left(\frac{p_j}{p_h}\right)^{\bar{\alpha}} \rho_{jh}(\bar{\alpha})(q')^{n-1} \qquad 3.2.1.13$$

where

$$q_1 = \max_{j \neq h}(q_{jh}') \quad \text{and}$$

$$\rho_{jh}(\bar{\alpha}) = \sum_{X_1} P_j(X_1)^{\bar{\alpha}} P_h(X_1)^{1-\bar{\alpha}}.$$

This approach gives the result of the theorem with q' defined differently from q and

$$A = C \sum_{h=1}^{m} \sum_{\substack{j=1 \\ j \neq h}}^{m} p_j^{\bar{\alpha}} p_h^{1-\bar{\alpha}} \rho_{jh}(\bar{\alpha}).$$

While 3.2.1.13 gives the strongest version of the theorem obtainable with this approach, there is no straightforward general algorithm for computing $\bar{\alpha}$, so that 3.2.1.9 is probably easier to use except when the distributions have a convenient form. Also, $\rho_{jh}(\bar{\alpha})$ is not necessarily algebraically less than $\rho_{jh}(\frac{1}{2}) = \rho_{jh}$, so 3.2.1.9 might give smaller values than 3.2.1.13 in some cases.

3.2.2 Bounds Based on the Distance Between Distributions

Several investigators [K-1, K-2, L-2] have used the Bhattacharyya coefficient to bound the probability of error of a maximum likelihood rule. Kadota and Shepp [K-1] and other statistical literature call this quantity the Hellinger integral. This coefficient is an inner product of two hypothesis conditional probability densities. Using ρ to denote the Bhattacharyya coefficient for two-class

hypothesis testing, the definition of ρ for discrete distributions is

$$\rho = \sum_{X} \left(P_1(X)P_2(X)\right)^{\frac{1}{2}}$$

where $P_i(X)$, $i = 1,2$, is the mass function under hypothesis H_i. From Schwarz's inequality one can see that $0 \leqslant \rho \leqslant 1$ since

$$\sum_{X} \left(P_1(X)P_2(X)\right)^{\frac{1}{2}} \leqslant \left(\sum_{X} P_1(X) \sum_{X} P_2(X)\right)^{\frac{1}{2}} = 1$$

and the arguments are non-negative over their domains. Several functions of ρ can be used to describe a "distance" between the density functions, with $-\log \rho$ being a favorite since it is non-negative and $-\log \rho = 0$ when $P_1(X) \equiv P_2(X)$. Since the Bhattacharyya coefficient has played such a significant role in the recent literature, several results will be presented in order to show its role in providing error bounds for m-class maximum likelihood hypothesis testing using dependent random variables.

In m-class hypothesis testing with $m \geqslant 2$, the inter-class Bhattacharyya coefficient, ρ_{ij}, is given by

$$\rho_{ij} = \sum_{X} \left(P_i(X)P_j(X)\right)^{\frac{1}{2}} \quad \text{for } i,j = 1,\ldots,m.$$

The derivations of error bounds related to the Bhattacharyya coefficient are designed by treating all of the observed patterns as a single pattern, rather than defining a bound that is a function of the number of patterns. This is in distinct contrast to the kinds of bounds derived by using either majority decision functions or equivocation. Further, it changes the mathematical treatment of the dependence to a matter of the computation of ρ, making that computation more complex as the number of observations increases.

In order to use the Bhattacharyya coefficient approach to bound the error probability after the k-th pattern in a sequence one must let the argument of $P_i(\cdot)$ be X^k, a sequence of real variables whose possible values are the set of possible sequences of the first k patterns. When the patterns have first order dependence, the probability $P_i(X^k)$ is given by

$$P_i(X^k) = \prod_{\gamma=1}^{k} P_i(X_\gamma | X_{\gamma-1})$$

where $P_i(X_1|X_0)$ is the prior probability of X_1 under hypothesis H_i, and $X^k = X_1, X_2, \ldots, X_k$. Then the interclass Bhattacharyya coefficient between class i and class j after k observations is given by

$$\rho_{ij}^{(k)} = \sum_{X^k} \left(\prod_{\gamma=1}^{k} P_i(X_\gamma | X_{\gamma-1}) P_j(X_\gamma | X_{\gamma-1}) \right)^{\frac{1}{2}} \qquad 3.2.2.1$$

When each pattern, X_γ, has n binary digits then X_γ has 2^n possible values, and the sum in 3.2.2.1 has 2^{nk} terms for each $\rho_{ij}^{(k)}$. Computing all $\rho_{ij}^{(k)}$ for $i < j$ requires summing $m(m-1)\cdot 2^{nk}$ terms. While this technique does not provide a closed form solution for the error bound as a function of the number of observations, it does suggest computerized experiments to determine a sequence of error bounds, subject to whatever limits exist on computing resources.

The computational technique for evaluating $\rho_{ij}^{(k)}$ can take a recursive form evidenced by expanding 3.2.2.1 as

$$\rho_{ij}^{(k)} = \sum_{X_k} \sum_{X_{k-1}} \left(P_i(X_k|X_{k-1})P_j(X_k|X_{k-1})\right)^{\frac{1}{2}} \cdot$$

$$\sum_{X_{k-2}} \left(P_i(X_{k-1}|X_{k-2})P_j(X_{k-1}|X_{k-2})\right)^{\frac{1}{2}} \ldots$$

$$\sum_{X_2} \left(P_i(X_3|X_2)P_j(X_3|X_2)\right)^{\frac{1}{2}} \cdot$$

$$\sum_{X_1} \left(P_i(X_2|X_1)P_j(X_2|X_1)P_i(X_1|X_0)P_j(X_1|X_0)\right)^{\frac{1}{2}}$$

3.2.2.2

Defining the factors in the equation above

$$r_{ij}(X_2) = \sum_{X_1} \left(P_i(X_2|X_1)P_j(X_2|X_1)P_i(X_1|X_0)P_j(X_1|X_0)\right)^{\frac{1}{2}}$$

and

$$r_{ij}(X_k) = \sum_{X_{k-1}} \left(P_i(X_k|X_{k-1})P_j(X_k|X_{k-1})\right)^{\frac{1}{2}} r_{ij}(X_{k-1})$$

for $k > 2$

gives

$$\rho_{ij}^{(k)} = \sum_{X_k} r_{ij}(X_k) \qquad \text{for } k \geq 2$$

With this technique, one can compute the $\rho_{ij}^{(k)}$ sequentially by saving the 2^n values of $r_{ij}(X_k)$ at each stage for use at the next stage.

The use of $\rho_{ij}^{(k)}$ in computing error bounds will now be considered. The quantity ρ_{ij} for a single observation appeared in Theorem 3.2.1.1 and it was pointed out that for the i.i.d. case the theorem became

$$E\left(H(\Theta|X^k)\right) \leq A\rho^{*k}$$

where $\rho^* \equiv \max_{i \neq j} \rho_{ij}$ and the superscript represents exponentiation. If one starts with the approach of the previous theorem and attempts to include $\rho_{ij}^{(k)}$ in the expression for the error bound, then one obtains the following theorem.

Theorem 3.2.2.1

Under the hypotheses of Theorem 3.2.1.1,

$$E\left(H(\Theta|X^k)\right) < C \sum_{h=1}^{m} \sum_{\substack{j=1 \\ j \neq h}}^{m} (p_j p_h)^{\frac{1}{2}} \rho_{hj}^{(k)}$$

Proof: The proof commences in the same manner as before through 3.2.1.4. Following 3.2.1.4,

$$E\left\{\prod_{i=1}^{k}\left[\frac{P_j(X_i|X_{i-1})}{P_h(X_i|X_{i-1})}\right]^{\frac{1}{2}}\middle|\Omega_h\right\} = \sum_{X^k}\prod_{i=1}^{k}\left[\frac{P_j(X_i|X_{i-1})}{P_h(X_i|X_{i-1})}\right]^{\frac{1}{2}}P_h(X^k)$$

$$= \sum_{X^k}\prod_{i=1}^{k}\left(P_j(X_i|X_{i-1})P_h(X_i|X_{i-1})\right)^{\frac{1}{2}} = \rho_{hj}(k)$$

so that

$$E\{H(\Theta|X^k)|\Omega_h\} < C\sum_{\substack{j=1\\ j\neq h}}^{m}\left(\frac{p_j}{p_h}\right)^{\frac{1}{2}}\rho_{hj}(k).$$

Therefore

$$E\left(H(\Theta|X^k)\right) < \sum_{h=1}^{m}p_h C\sum_{\substack{j=1\\ j\neq h}}^{m}\left(\frac{p_j}{p_h}\right)^{\frac{1}{2}}\rho_{hj}(k)$$

which is the result stated in the theorem.

Q.E.D.

While this does show the role of the interclass Bhattacharyya coefficient for k observations, it is not as tight a bound as Lainiotis [L-3] has achieved working from Chu and Chueh's premise that

$$P_e < \sum_{i<j} P_e(i,j) \quad .$$

Lainiotis' upper bound on the error probability of the maximum likelihood decision is given in the following theorem.

Theorem 3.2.2.2

In the m-class hypothesis testing problem with prior probability p_i of hypothesis H_i and conditional probability $P_i(X)$ under hypothesis H_i, the probability of error of the maximum likelihood decision is bounded above by

$$P_e \leq \sum_{i<j} p_i^{\alpha_{ij}} p_j^{1-\alpha_{ij}} \sum_X P_i(X)^{\alpha_{ij}} P_j(X)^{1-\alpha_{ij}} \qquad 3.2.2.3$$

for $0 \leq \alpha_{ij} \leq 1$ and $i,j = 1,\ldots,m$.

Proof: The proof is given by Lainoitis [L-3] and will not be repeated here.

For maximum likelihood decisions using the Bayes posterior distribution after k observations on dependent random variables, the argument X in 3.2.2.3 becomes X^k, and the $P_i(X^k)$ are the Bayes posterior distributions.

While there is no general way to choose the α_{ij} in 3.2.2.3 to minimize the bound, one is free to choose any convenient value in [0,1] for α_{ij}. When the hypotheses have equal prior probability, choosing $\alpha_{ij} = \frac{1}{2}$ for all i and j, and applying Lainiotis' theorem gives

$$P_e \leq \frac{1}{m} \sum_{i<j} \rho_{ij}$$

for a single observation or

$$P_e \leq \frac{1}{m} \sum_{i<j} \rho_{ij}^{(k)} \qquad 3.2.2.4$$

for the decision after k observations, which compares with $P_e < C(m-1)\rho_{ij}^{(k)}$ in Theorem 3.2.2.1. Even using 3.2.2.4, there is the possibility that the bound is greater than 1. There are $m(m-1)/2$ terms in the sum in 3.2.2.4, and if all of the ρ_{ij} or $\rho_{ij}^{(k)}$ are very close to 1, then the bound can be close to $(m-1)/2$. However, in the case described, all of the distributions are nearly identical, and one would anticipate difficulty in making decisions.

Lainiotis' theorem agrees with results published by Kailath [K-2] and Kadota and Shepp [K-1]. Kailath's upper and lower bounds for the two class, equal prior probability, maximum likelihood decision are

$$\frac{1}{4}\rho^2 \leq P_e \leq \frac{1}{2}\rho$$

while Kadota and Shepp retain the prior probabilities, p_1 and p_2, for the bounds

$$\frac{1}{2}\min(p_1,p_2)\rho^2 \leq P_e \leq \sqrt{p_1 p_2}\,\rho.$$

G. T. Toussaint, in a recently published paper [T-3], derived an upper bound on the interclass Bhattacharyya coefficient by using the Kolmogorov variational distance, which might be easier to calculate than is the Bhattacharyya coefficient. The Kolmogorov variational distance between two hypothesis conditional distributions is defined as

$$V_{ij} = \sum_X |P_i(X) - P_j(X)| .$$

The theorem is as follows:

Theorem 3.2.2.3

(Toussaint) In m-class hypothesis testing using maximum likelihood decisions with equal prior probabilities, the probability of error is bounded by

$$P_e \leq \frac{1}{m} \sum_{i<j} (1 - V_{ij}/2)$$

Proof: Using Chu and Chueh's upper bound

$$P_e \leq \sum_{i<j} P_e(i,j)$$

with

$$P_e(i,j) = \sum_X \min\{p_i P_i(X); p_j P_j(X)\} \qquad 3.2.2.5$$

and

$$p_i = p_j = \frac{1}{m} \text{ for } i,j = 1,\ldots,m$$

gives

$$P_e \leq \frac{1}{m} \sum_{i<j} \sum_X \min\{P_i(X);P_j(X)\}. \qquad 3.2.2.6$$

Kailath [K-2] shows that

$$P_e(i,j) = p_i - \frac{1}{2} \sum_X |p_i P_i(X) - p_j P_j(X)| \qquad 3.2.2.7$$

One can use 3.2.2.5 on the left of 3.2.2.7, so that for $p_i = p_j = \frac{1}{m}$, the prior probabilities factor and cancel leaving 3.2.2.7 reduced to

$$\min\{P_i(X);P_j(X)\} = 1 - \frac{1}{2} \sum_X |P_i(X) - P_j(X)|$$

$$= 1 - \frac{1}{2} V_{ij} \qquad 3.2.2.8$$

Using 3.2.2.8 in 3.2.2.6 completes the proof.

Q.E.D.

The functional form of the Kolmogorov variational distance prohibits the type of factoring used in 3.2.2.2 to compute $\rho_{ij}(k)$. So when the argument of the hypothesis conditional probability is X^k, the sequence of k observations, the storage requirements for $P_i(X^k)$ grow exponentially with k.

3.3 Summary

The majority decision procedure is designed to allow one to apply the Central Limit Theorem to describe the error probability as the number of observations grows without bound. The behavior of the resulting normal distribution function is well known, and tabulated values are readily available if one wants to obtain an estimate of the error probability in a specific case. There is no way to compute a number of observations (finite) which is sufficient to guarantee that such an estimate is an upper bound on the error probability. Computationally, this procedure has the advantage that, once the single pattern variational distance is computed, arguments for the limiting distribution can be easily hand calculated.

Consideration of the equivocation of a sequence of observations provides the valuable insight that the expectation of the amount of additional information that can be obtained by taking additional samples goes to zero as the number of samples increases. The functional form of the equivocation, coupled with the probability of error expression for the optimum processor described in Chapter II, defines a computable upper bound for the error probability--which was missing in the first approach using the Central Limit Theorem. Once the parameters of the bound have been computed--generally a job which requires the aid of a digital computer--one can use a slide rule to

determine the number of observations required in order for the error bound to assure a given error level.

The error bounds obtained from the Bhattacharyya coefficient and the variational distance appear to be easier to calculate than the equivocation bound, but have the drawback that the bound might exceed 1 regardless of the number of observations. Also the computation of $\rho_{ij}^{(k)}$ and V_{ij}, when the distance is between $P_i(X^k)$ and $P_j(X^k)$, can be as burdensome as computing the quantities needed for the equivocation bound. The recursive method described for the Bhattacharyya coefficient produces an algorithm for which the computation increases linearily with the number of observations, and the storage requirements are fixed. The variational distance bound, on the other hand, forces exponentially growing amounts of computation and storage as one tries to compute the bound for successively larger numbers of observations.

CHAPTER IV

EXAMPLES OF DECISION PROCESSES AND ERROR BOUNDS

The results presented in the previous two chapters suggest a wide variety of test cases to illustrate the theory involved. Several such tests have been simulated on the CDC 6500 computer, and the results are summarized in this chapter. The first example shows how the processor learns the symbol code length of the source when the sources have different code lengths and there is no channel noise. Next, a noisy channel--the binary symmetric channel--is used, and the sources all have the same code length. In this latter case, only the symbol distributions distinguish the sources, and the channel error rate is unknown. Decisions are compared with decisions made using a minimum distance estimator. Computational overhead for computing Bayes posterior conditional distributions from the 1-dependent observations is examined, and a processing technique is proposed which reduces the computational overhead to the amount one would require if the observations were independent. The proposed processing

technique is used on the binary symmetric channel with favorable experimental results. The convergence and error probability theorems of Chapters II and III do not apply to this suboptimum technique.

Finally, the various error bounds presented in Chapter III are compared for a specific example of distributions for three sources. The upper bounds in particular indicate a requirement for a large number of observations to obtain a desired error rate, although the limiting distribution derived from the majority decision function suggests that low error probability could be obtained from relatively fewer observations.

4.1 The Bayes Decision Process Applied to the Symbol Synchronization Problem

Experiments programmed for the CDC 6500 computer illustrate the behavior of the posterior distributions used in the optimum decision procedure described in Chapter II. Figure 4.1 shows the results of one set of experiments. In these experiments each of three sources used a different symbol code length--one, two and three binary digits, respectively. The probability law governing the i.i.d. symbol selection process for the respective sources is shown in Table 4.1. The receiver was simulated by making observations on patterns containing 3 binary digits, the smallest pattern length that could be used and still satisfy the requirement that the length

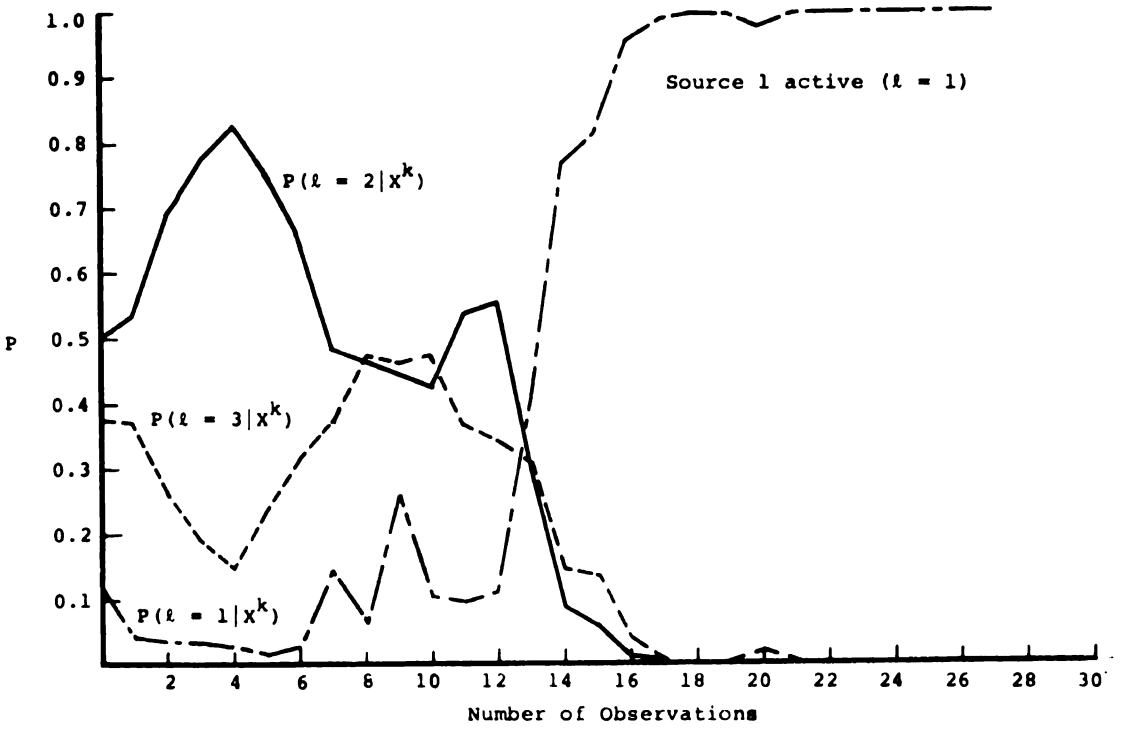


Figure 4.1 (a) Posterior Distribution of the Source Index Parameter

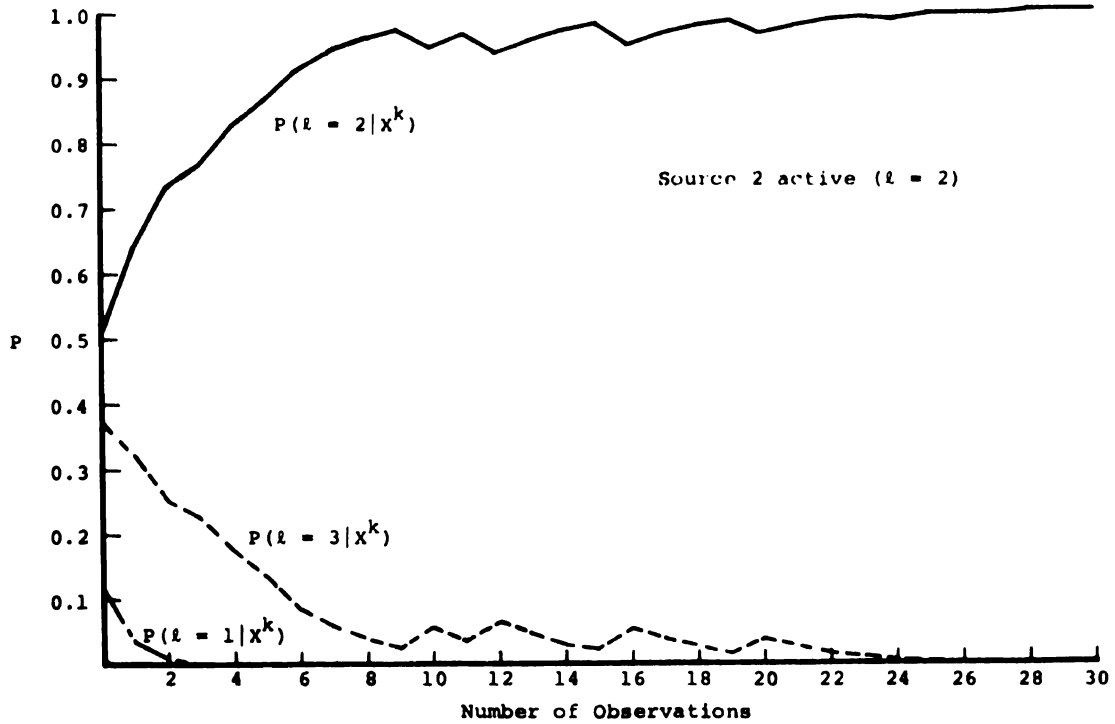


Figure 4.1 (b) Posterior Distribution of the Source Index Parameter

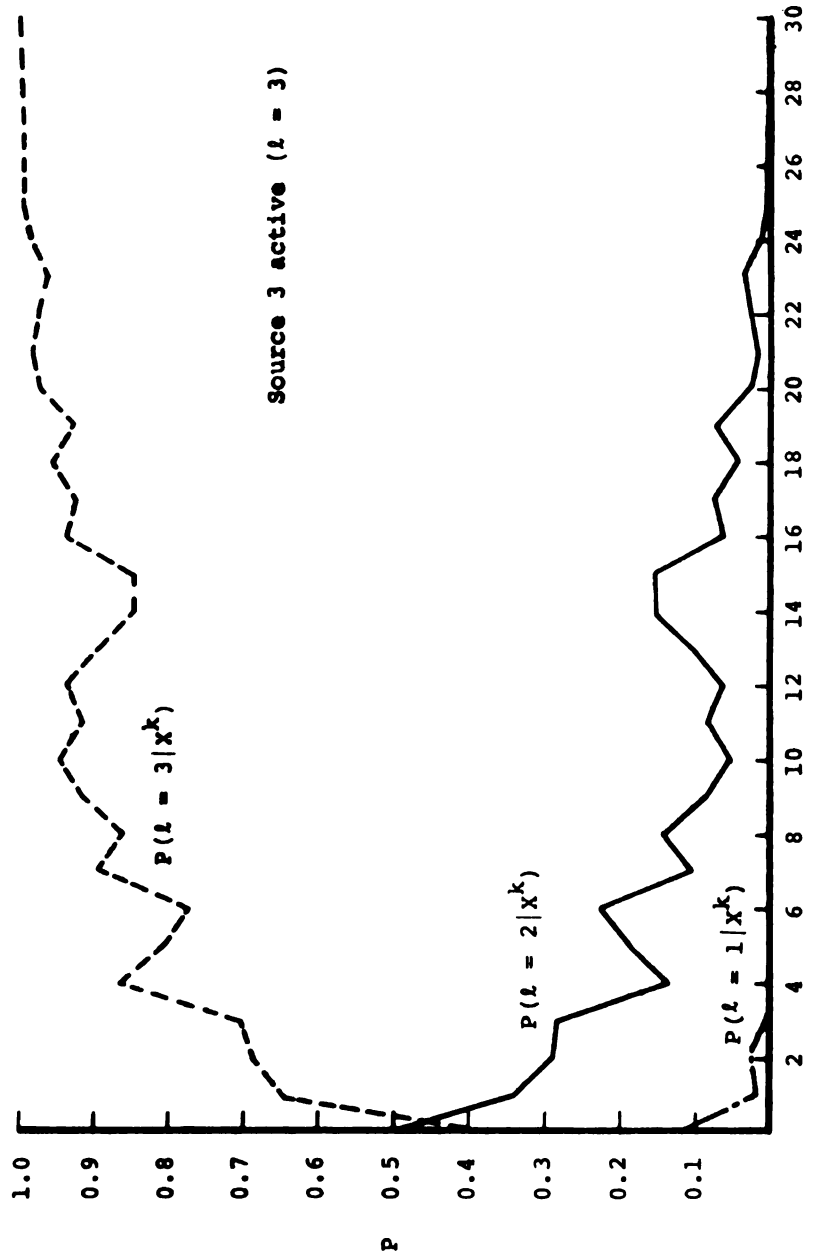


Figure 4.1 (c) Posterior Distribution of the Source Index Parameter

Table 4.1 Symbol Generating Probabilities.

Source Number ℓ	Source Code Length m_ℓ	Source Code	Probability That Code is Generated
1	1	0	0.75
		1	0.25
2	2	00	0.127
		01	0.375
		10	0.375
		11	0.125
3	3	000	0.09375
		001	0.125
		010	0.15625
		011	0.125
		100	0.125
		101	0.09375
		110	0.09375
		111	0.1875

Symbol selections by the sources mentioned in the example of Figure 4.1 are governed by the respective distribution shown above.

of X_k be at least the longest symbol code length. Figure 4.1 shows the posterior conditional probability that each source is active as a function of the number of observations. The ordinate intersections represent the prior probabilities. After any particular observation, the decision procedure decides that the source with the greatest posterior probability is active.

Figure 4.1 illustrates that the initial bias, imposed by the prior distributions, can be overcome by this decision process. The discussion in Chapter III suggests that the distance between the probability distributions of the observations, which is influenced by the source distributions, should affect the number of observations required to overcome this initial bias. Posterior distributions for the synchronization instant $P(T_k|\ell,X^k)$ were computed as well. In the first example with $m_\ell = 1$, T_k is 1 for all k, so that once the source is correctly decided the synchronization instants are obvious. In the second example with $\ell = 2$ and $m_\ell = 2$, the true value of T_k alternates between 1 and 2. Since the prior probability for the correct value of T_1 was 0.714, the posterior probabilities $P(T_k|\ell = 2,X^k)$ produced correct decisions for all k. In the third example with $\ell = 3$ and $m_\ell = 3$, the prior probability for the correct value of T_1 was 0.222, vs. values of 0.333 and 0.445 for the other possibilities for T_1. This, combined with the probability

distribution for source 3, led to needing over 20 pattern observations before correct synchronization decisions were made, even though the source decision was correct after every observation.

4.2 The Binary Symmetric Channel

The binary symmetric channel illustrates the notion of an unknown channel parameter. As the name implies, the noise in the channel has the net effect of changing a code digit one to a digit zero with probability p and of changing a code digit zero to a one with probability p, as diagrammed in Figure 4.2. Thus p is the probability that a binary digit is complemented as it passes through the channel, and will be called the complementation rate.

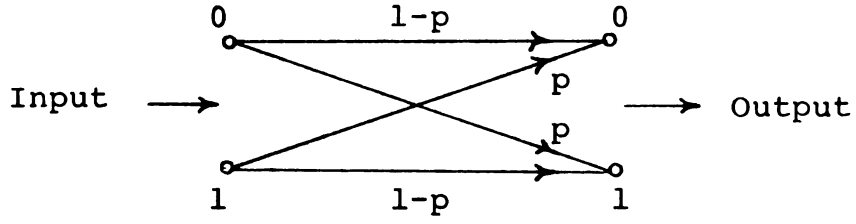


Figure 4.2

The Binary Symmetric Channel

A digit passes through the channel unchanged with probability 1-p. As in the data generation model previously considered, the channel is connected to a source which generates a binary digit code of length m_ℓ, and transmits the codes in a continuous stream. The receiver processes

the output as n binary digit patterns, X_k, k = 1,2,..., $n \geq m_\ell$ for all ℓ. Y_k will denote the input which produced the output X_k. Both X_k and Y_k can be any one of the 2^n vectors having n binary elements, but X_k is not necessarily equal to Y_k because of the noise properties of the channel. The quantity p is an unknown parameter.

The full set of assumptions for this special case follows.

1. Patterns of n binary elements, X_k, are received through the binary symmetric channel of Figure 4.2.

2. The value of the parameter p is unknown.

3. Changes of distinct binary digits in a pattern are mutually independent.

4. Given the index of the active source, ℓ, and synchronization instant, T_k, the distribution of Y_k at the input, denoted by $P_{in}(Y_k|\ell,T_k)$, is known.

Define $C(Y_k,X_k)$ as the number of bit positions in which Y_k and X_k differ; $C(Y_k,X_k)$ is the Hamming distance between Y_k and X_k.

Let Z_j, $j\varepsilon\{1,2,\ldots,2^n\}$, denote the base 10 values corresponding to the strings of base 2 digits which can be assumed by Y_k and X_k. The probability of receiving $X_k = Z_j$ when $Y_k = Z_i$ was transmitted is

$$P(X_k = Z_j | Y_k = Z_i) = p^{C(Z_i,Z_j)} (1-p)^{n-C(Z_i,Z_j)}$$

$$i,j \in \{1,2,\ldots,2^n\}$$

The conditional probability of receiving a particular vector, $X_k = Z_j$, can be represented by

$$\begin{aligned} & P(X_k = Z_j | p,\ell,T_k) \\ &= \sum_{i=1}^{2^n} P_{in}(Y_k = Z_i | p,\ell,T_k) P(X_k = Z_j | Y_k = Z_i, p,\ell,T_k) \\ &= \sum_{i=1}^{2^n} P_{in}(Y_k = Z_i | \ell,T_k) P(X_k = Z_j | Y_k = Z_i) \\ &= \sum_{i=1}^{2^n} P_{in}(Y_k = Z_i | \ell,T_k) p^{C(Z_i,Z_j)} (1-p)^{n-C(Z_i,Z_j)} \\ &\triangleq Q_j(p,\ell,T_k) \end{aligned}$$

$$i,j \in \{1,2,\ldots,2^n\},\ 0 \leq p \leq 1,\ \ell = 1,\ldots,L,$$

and $T_k = 1,\ldots,\ell$.

This equation defines the quantity $Q_j(p,\ell,T_k) = P(X_k = Z_j | p,\ell,T_k)$ as a polynomial in p of degree n. Given ℓ and T_k, there are 2^n of these polynomials, one for each possible value of X_k.

Theorem 2.5.1 says that if members of the family $\{P(X_k | p,\ell,T_k)\}_{\ell,T_k}$ are distinct, then strongly consistent estimators exist for p, ℓ, and T_k. Saying that members

of the family are distinct means that for any $(a,b)\neq(c,d)$ there exists at least one value Z_j such that $P(X_k = Z_j|p,\ell = a,T_k = b)\neq P(X_k = Z_j|p,\ell = c,T_k = d)$ when those quantities are defined. As a result, the hypotheses of Theorem 2.5.1 require that no two synchronizations have exactly the same set of 2^n polynomials $Q_j(p,\ell,T_k)$ for all 2^n values of j. Equivalently, for $(a,b)\neq(c,d)$ $Q_j(p,\ell = a,T_k = b)$ must not have the same set of coefficients as $Q_j(p,\ell = c,T_k = d)$, $j\varepsilon\{1,2,\ldots,2^n\}$, when (a,b) and (c,d) are such that $Q_j(\cdot)$ is defined. In turn, the coefficients of $Q_j(p,\ell,T_k)$ are defined above to be linear combinations of $P_{in}(Y_k|\ell,T_k)$, so it is reasonable to look for a condition on the family $\{P_{in}(Y_k|\ell,T_k)\}_{\ell,T_k}$ which guarantees that linear combinations of members of that family are unique.

Identifiability is such a sufficient condition. Consequently, one can proceed with the Bayes decision process to simultaneously decide the source, synchronization, and channel complementation rate and assume that the sequences of decisions will converge provided that $\{P_{in}(Y_k|\ell,T_k)\}_{\ell,T_k}$ is known to be an identifiable family [T-1, T-2, Y-3] or provided that the members can be shown to be linearly independent.

The learning capability of the processor was demonstrated for a binary symmetric channel with unknown complementation rate. The complementation rate was the parameter

to be learned. A Fortran language program, BSC MC--for Binary Symmetric Channel, Monte Carlo--provides appropriate data generation, channel simulation, and Bayes rule processing. For coding convenience, BSC MC processes sources whose code length is 3 binary digits and uses observations of 3 binary digits. BSC MC allows up to 10 discrete values for the channel complementation rate; two values, 0.05 vs. 0.1, were used in the examples which are reported here. The program is set to cut off after 500 observations. Then the prior distributions can be re-initialized for as many Monte Carlo iterations as one desires.

As a check on the data generated, BSC MC maintains counts of the number of times each symbol is generated, the number of bits changed by the channel and the number of times each observation vector value is observed. This last is used to compute the empirical mass function $\overline{P}(X_k)$ which can be used in an alternative minimum distance estimate in which

$$\text{distance} = \min_{\ell, T_k} \max_{X} \left| \overline{P}(X_k) - P(X) \right| .$$

The decision procedure decides that the values of the minimization arguments for which the minimum is achieved give the source and synchronization. In the examples run, the optimum decision and the minimum distance decision

agreed with respect to the source and synchronization, but not always with respect to the channel parameter (channel complementation rate).

A second program, called BMCIND (for Binary symmetric channel, Monte Carlo, INDependent) is identical to BSC MC except for a subroutine named POST which computes the posterior distributions. For BMCIND, POST computes the posterior distributions as if the observations were independent. After the k-th observation, the version of POST used with BMCIND computes

$$P(p|\ell,T_k,X_k) = \frac{P(X_k|p,\ell,T_k)P(p|\ell,T_k,X_{k-1})}{\sum_p P(X_k|p,\ell,T_k)P(p|\ell,T_k,X_{k-1})}$$

where the denominator is defined to be $P(X_k|\ell,T_k)$,

$$P(T_k|\ell,X_k) = \frac{P(X_k|\ell,T_k)P(T_k|\ell,X_{k-1})}{\sum_{T_k} P(X_k|\ell,T_k)P(T_k|\ell,X_{k-1})}$$

where the denominator is defined to be $P(X_k|\ell)$,

$$P(\ell|X_k) = \frac{P(X_k|\ell)P(\ell|X_{k-1})}{P(X_k|\ell)P(\ell|X_{k-1})} .$$

By contrast, the version of POST used in connection with BSCMC computes

$$P(p|\ell,T_k,X_k) = \frac{P(X_k|p,\ell,T_k,X_{k-1})P(p|\ell,T_k,X_{k-1})}{\sum_p P(X_k|p,\ell,T_k,X_{k-1})P(p|\ell,T_k,X_{k-1})}$$

where the denominator is defined to be $P(X_k|\ell,T_k,X_{k-1})$,

$$P(T_k|\ell,X_k) = \frac{P(X_k|\ell,T_k,X_{k-1})P(T_k|\ell,X_{k-1})}{\sum_{T_k} P(X_k|\ell,T_k,X_{k-1})P(T_k|\ell,X_{k-1})}$$

where the denominator is $P(X_k|\ell,X_{k-1})$, and

$$P(\ell|X_k) = \frac{P(X_k|\ell,X_{k-1})P(\ell|X_{k-1})}{\sum_{\ell} P(X_k|\ell,X_{k-1})P(\ell|X_{k-1})} .$$

The "posterior distributions," computed in connection with BMCIND, when used with a maximum likelihood decision procedure, learned the unknowns but appeared to converge less rapidly than did the Bayes processor. The quantities computed by the version of POST used with BMCIND coincided with the Bayes posterior conditional distributions only if the observations were independent. The motivation for using these quantities when the observations are dependent stems from a consideration of storage requirements and computation volume, which is discussed next.

In considering the storage requirements for computing the Bayes posterior conditional distributions for the processor described in Chapter II, one must first realize that there are 2^n values of the conditional probability $P(X_k|\ell,T_k,X^{k-1})$ for each set of values of the conditional arguments. The dependence on X^{k-1} is a result of the possibility that $T_k \neq 1$, i.e., a symbol code begins in X_{k-1}

and ends in X_k as described in Section 2.1. Since successive symbol codes are independent, only the cases in which symbol codes overlap observations X_{k-1} and X_k will tend to affect the storage of the posterior conditional distributions. If $T_k = 1$, then the conditional probability is independent of X_{k-1}, and

$$P(X_k|\ell,T_k,X^{k-1}) = P(X_k|\ell,T_k).$$

If $T_k = 2$, then the conditional probability is dependent on the last $m_\ell - 1$ digits of X_{k-1}. In general, $P(X_k|\ell,T_k,X^{k-1})$ is dependent on the last

$$d = [m_\ell - T_k + 1]_{\text{mod } m_\ell}$$

digits of X_{k-1}. While d is in fact a function of m_ℓ and T_k, writing d rather than $d(m_\ell,T_k)$ is convenient because d will be used as an exponent in subsequent expressions.

Taking into account the above description of how the value of T_k describes the dependence on X_{k-1}, one can see that there are 2^{n+d} values of $P(X_k|\ell,T_k,X^{k-1})$ for each value of ℓ and T_k. Given $\ell = \lambda$, there are m_λ values of T_k, so for each source, λ, there are

$$\sum_{T_k=1}^{m_\lambda} 2^{n+d} = 2^{n+0} + 2^{n+m_\lambda-1} + \ldots + 2^{n+1}$$

$$= 2^n \sum_{i=0}^{m_\lambda-1} 2^i = 2^n(2^{m_\lambda} - 1)$$

values of $P(X_k|\ell,T_k,X^{k-1})$ to compute and store, which could tax the capacity of a very large computer. For example, suppose one of the sources which might be observed uses a symbol code length, including parity digits, of 9 binary digits. Suppose further that this is the longest possible source code length, so n is chosen to be 9. Then

$$\sum_{T_k=1}^{9} 2^{n+d} = 2^9(2^9 - 1) = 2^{18} - 2^9 > 2^{17}.$$

This shows that just storing the conditional distribution for a single source could use up all of the immediately accessible, individually addressable core storage of the latest model computers. The usual storage requirement for a Markov chain having 2^9 states is 2^{18} values, but the analysis based on the value of T_k was aimed at identifying redundant values that need not be stored.

The technique used by subroutine POST in the BMCIND program would reduce the storage requirements to 2^9 in the case cited above, or a reduction factor of over $2^{(n-1)}$.

Table 4.2 summarizes the results of one computer experiment with the binary symmetric channel. In this instance, each of three sources used three binary digits in its symbol codes. The channel error rate was to be either 0.05 or 0.10. Ten runs were made, and after the 500-th observation a decision was made about the channel

error rate, about which source was active, and about the synchronization instant. Results of the optimum processor (BSC MC), one suboptimum processor (BMCIND), and the minimum distance decision described earlier in this section are compared.

Table 4.2 Comparison of Three Decision Processes.

	Number of Correct Decisions in 10 Monte Carlo Runs		
	Channel Bit Complementation Rate	Active Source	Synchronization Instant
BSC MC	4	10	10
BMCIND	2	10	10
Minimum Distance	6	10	10

The Bayes decision process, a Bayes-like process, and a minimum distance decision process were used to decide the bit complementation rate, source, and synchronization.

In these experiments, the prior distributions were randomly generated before the first run, and the same prior distributions were re-established for each of the following 9 runs.

It is heartening to note that BSC MC, which uses the Bayes posterior distribution, performs somewhat better than BMCIND, which uses the Bayes formula with the marginals of the conditional distributions. But the superior performance of the minimum distance estimator was unexpected.

Examination of the history of the posterior distributions computed by both BSC MC and BMCIND shows that the decisions made after the earlier observations vary, but after about 25 observations, all subsequent decisions are identical, whether they are correct or not. This suggests that a type of bias develops that is unlikely to be overruled if the probability distributions are very close. The posterior distributions of quantities that are quite distinctive, such as the source and synchronization instant in this case, tend to be capable of producing correct decisions even if the decisions regarding other quantities are incorrect.

4.3 Error Estimates and Bounds

This section presents numerical results obtained by applying the various error bounds of Chapter III to a specific example. Details of the example were chosen for computational expediency rather than to represent a particular application. Specifically, the example is not typical of the models one would expect in the symbol synchronization problem.

In the example there are three pattern classes, for which a pattern is one binary digit. Each class has a stationary first order dependent distribution. This is intended to mean that for fixed ℓ and m--ℓ,mε{0,1}--and any class i

$P_i(X_k = \ell | X_{k-1} = m)$ is constant for all k = 2,3,....

The values used for the class conditional joint probability distributions are tabulated below.

	i=1	i=2	i=3
$P_i(X_k = 0, X_{k-1} = 0)$	0.40	0.30	0.10
$P_i(X_k = 0, X_{k-1} = 1)$	0.15	0.05	0.20
$P_i(X_k = 1, X_{k-1} = 0)$	0.20	0.45	0.30
$P_i(X_k = 1, X_{k-1} = 1)$	0.25	0.20	0.40

The values of the joint distributions lead to the sample conditional distributions below, which are used to obtain the upper bound based on the equivocation measure.

	i=1	i=2	i=3
$P_i(X_k = 0 \mid X_{k-1} = 0)$	0.667	0.400	0.250
$P_i(X_k = 0 \mid X_{k-1} = 1)$	0.375	0.200	0.333
$P_i(X_k = 1 \mid X_{k-1} = 0)$	0.333	0.600	0.750
$P_i(X_k = 1 \mid X_{k-1} = 1$	0.625	0.800	0.667

An important part of the equivocation bound argument, 3.2.1.6 of Chapter III, states that

$$E\left\{\left[\frac{P_j(X_k|X_{k-1})}{P_h(X_k|X_{k-1})}\right]^{\frac{1}{2}}\middle|\,\Omega_h\right\} < 1 \text{ for } j\neq h$$

A matrix of the values obtained for the above expectations obtained for this example is given below.

	h=1	h=2	h=3
j=1	1.000	.968	.963
j=2	.971	1.000	.988
j=3	.945	.987	1.000

The diagonal terms should obviously be 1 and serve as a check on the computation.

The classes were assumed to be equally probable. The prior class conditional distribution of the first observation and the corresponding interclass Bhattacharyya coefficients are tabulated below. The values of $P_i(X_1 = \cdot)$ were chosen to make $P_i(X_k)$ stationary with respect to k.

	i=1	i=2	i=3
$P_i(X_1 = 0)$	0.529	0.25	0.308
$P_i(X_1 = 1)$	0.471	0.75	0.692

Prior Class Conditional Distributions
of the First Observation

$$\rho_{12} = \rho_{21} = 0.958$$

$$\rho_{13} = \rho_{31} = 0.974$$

$$\rho_{23} = \rho_{32} = 0.997$$

Interclass Bhattacharyya Coefficients for the Distribution of the First Observation

Assuming that C of Lemma 1, Chapter III, is not greater than 2, the value of A in 3.2.1.10, is 3.907, and the value of q in 3.2.1.9 is 0.988. So the error bound is

$$P_e < 1.953 \times 0.988^{k-1}.$$

Under this approach, one is very much at the mercy of the value of q, which in this case is very close to the maximum possible value, one. Figure 4.3 shows that for this example the equivocation bound gives the largest values of all the techniques which were compared. Approximately 367 observations would be required in order for this error bound to assure an error rate of less than 0.05.

Comparing the above analysis based on the equivocation bound with the asymptotic distribution from the majority decision function approach reveals sizable differences in the error estimates. Using the form for the m pattern class case, one calculates

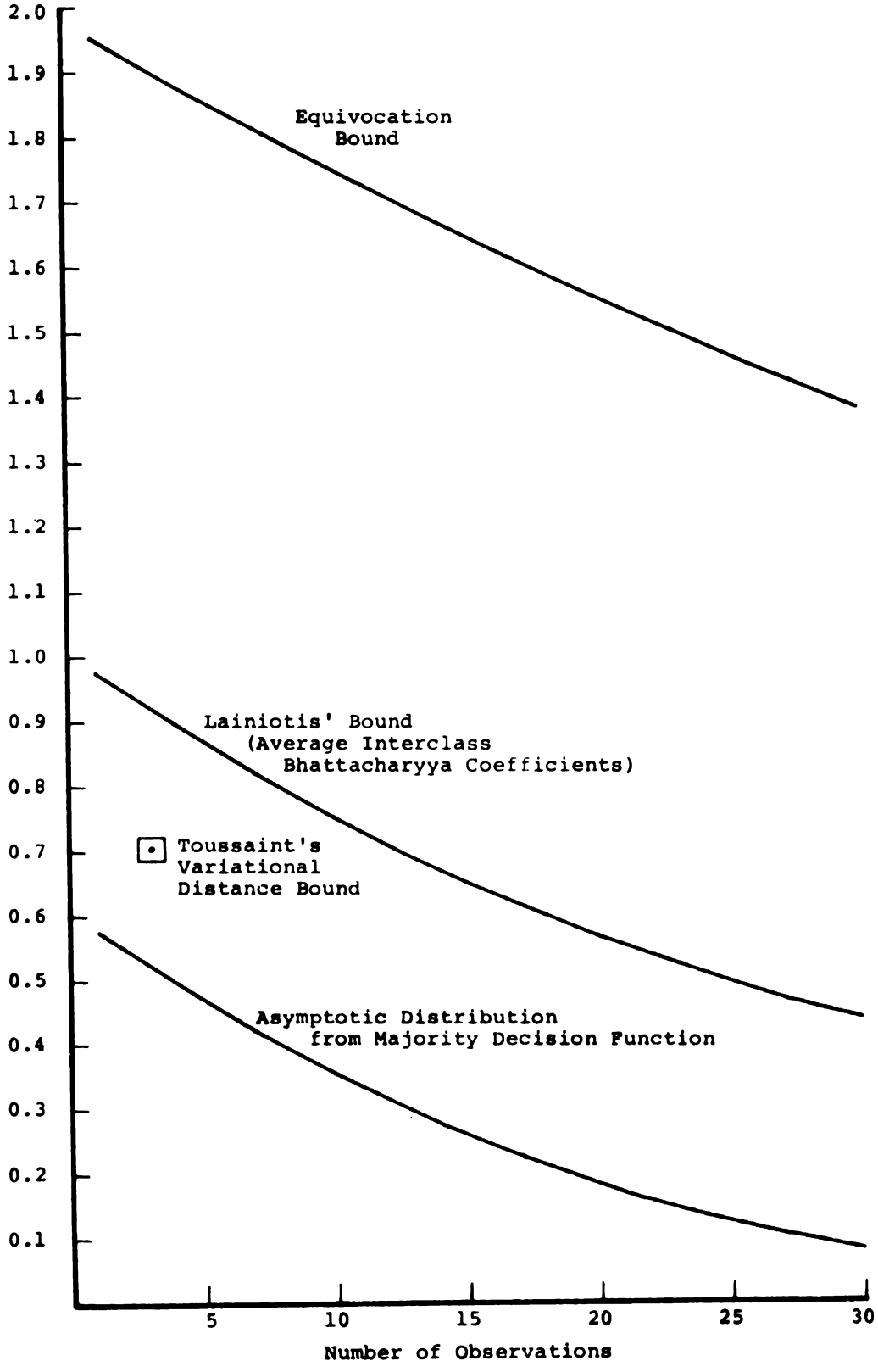


Figure 4.3 Comparison of Error Bounds

$$P_{\hat{i}}(X_k) = \frac{1}{1 - p_i} \sum_{\substack{j=1 \\ j \neq i}}^{m} p_j P_j(X_k)$$

and obtains the values tabulated below.

	i=1	i=2	i=3
$P_{\hat{i}}(X_k = 0)$	0.279	0.419	0.390
$P_{\hat{i}}(X_k = 1)$	0.721	0.581	0.610

The next task is to find δ such that

$$\sum_{X_k} |P_i(X_k) - P_{\hat{i}}(X_k)| \geq 2\delta$$

where the Kolmogorov variational distances on the left turn out to be 0.501, 0.337, and 0.164, respectively, for i = 1,2,3. Taking one-half the minimum distance for δ gives $\delta = 0.082$, and $\varepsilon = \frac{1}{2} - \frac{\delta}{4} = 0.479$. The values of the asymptotic distribution of the error probability as a function of the number of observations is given in the lowest curve of Figure 4.3 labeled "Majority Decision Function Asymptotic Distribution."

The bound suggested by Lainiotis in 3.2.2.4 is plotted as the middle curve in Figure 4.3. The values are all smaller than the values obtained from the equivocation method, plotted in the top curve, for the number of observations investigated. Whether this is true for large numbers of observations is not clear, since

the form of the Lainiotis bound does not reveal its asymptotic behavior. A consequent advantage of the equivocation bound is that, once the coefficient and base are calculated, one can estimate values for large numbers of samples by using a slide rule. Conversely, the Lainiotis bound requires a rather tedious recalculation, best done on a digital computer, for each successive number of observations.

As one comes to the error bound proposed by Toussaint, the storage requirements for the probability distributions threaten to be very costly. However, the bound was computed on the basis of three observations and came out to be 0.707, the smallest value provided by any of the analyses. The value is spotted on Figure 4.3.

CHAPTER V

CONCLUSIONS AND RECOMMENDATIONS

5.1 Summary of the Thesis

Pattern recognition techniques offer a choice between unsupervised learning, which requires prior knowledge of the probability distributions governing the events producing the patterns, and supervised learning, which requires a set of training data for computing parameters of the decision function. This thesis has ignored supervised learning techniques and the problems that arise in determining that the training data are sufficiently representative of the total population to assure a low probability of error. Instead, this work has demonstrated that certain statements about probability of error for m-class ($m > 2$) unsupervised pattern recognition can be extended, in an appropriately modified form, to problems involving statistical dependence.

The motivating problem for this research, symbol synchronization for an unknown source code length, has been shown to be a problem in unsupervised learning with statistically dependent data. Several solutions to this problem have been presented, including a Bayes decision

process and a stochastic approximation technique, both of which were shown to produce a sequence of decisions which converge to the correct decision. A suboptimum technique, formally like the Bayes process for independent random variables, was suggested, and empirical results were encouraging.

New results include applying the interclass Bhattacharyya coefficient to a sequence of observations having first order dependence in order to bound the probability of error of the Bayes decision process. Other error bounds are presented, one based on the expectation of the additional information in successive samples and one based on the Kolmogorov variational distance. The asymptotic distribution of the error probability for a suboptimum process provides an in-the-limit statement about the error rate of the Bayes decision procedure. It is interesting to note that the Bhattacharyya coefficient appears in the information theoretic bound and the Kolmogorov variational distance plays a role in the asymptotic distribution of the error probability. The probability of error statements which are applied to decision problems using dependent random variables have turned out to use these two measures on the distributions, the Central Limit Theorem and expectations of products of a very limited class of dependent random variables. In all of this, the computational expediency of the resulting algorithms has been a foremost consideration.

5.2 Recommendations for Continued Research

Unsupervised learning methodology would benefit from an influx of new decision processes which could be used as alternatives to the Bayes decision process. The majority decision techniques may be an important step in this direction. Chu and Chueh recommended that the error estimates derived from the majority decision process be considered as upper bounds on the error of the Bayes process, but that decisions should be made by the Bayes process in order to achieve the minimum decision error probability. An alternative approach would be to use easily computed local decisions as each pattern is observed and refine the estimate of the error probability based on properties of those local decision functions. This is essentially the philosophy behind the Bayes majority decision function, and others might be feasible. The objective, of course, is to reduce the computational burden imposed by the Bayes decision process but with a technique whose convergence and error properties could be well defined. The "posterior distributions" described in Chapter IV in connection with the program called BMCIND provided a computational advantage to the Bayes posterior distributions, but the convergence properties of the BMCIND process need to be determined. The BMCIND process was a semi-Bayes process which would be the Bayes decision process if the observations were independent. As work moves toward decision processes which are more and more unlike the Bayes decision process, it will most likely be

conducted by considering i.i.d. random variables at first, with the hope of generalizing to dependent random variables at a later time.

Since the Bhattacharyya coefficient has proved to have an important relationship to the error probability for the Bayes decision process, additional studies of its properties might prove useful. In particular one might be able to describe the class of distributions for which $\rho_{ij}(k)$ defined by 3.2.2.1 is a monotonically decreasing function of k. One could also attempt to determine relationships between the Bhattacharyya coefficient and error rates for non-Bayes decision procedures in either a supervised or an unsupervised mode.

Supervised learning techniques include the use of a linear combination of functions called potential functions [A-2, B-1, P-4] to approximate unknown probability distributions. Unsupervised learning has traditionally proceeded by assuming that, if one did not have classified training data, then one would assume that the probability distributions for the pattern classes are known to provide a starting point around which a decision process can be built. Knowing the probability distributions has been a demanding assumption for unsupervised learning. It is tempting to try to relax that assumption, and perhaps that could be done by using potential functions. The first task would be to determine whether there are

any conditions for the traditional unsupervised learning problem (in which all pattern classes are represented randomly in the data) under which combinations of potential functions would provide useful estimates of unknown pattern class distributions.

In general the properties of supervised learning techniques using dependent observations are not described in existing literature. The work of C. K. Chow is an exception, of course. Aside from the mathematical complexities introduced by considering dependent random variables, there are problems in that a large number of dependency models are candidates for consideration. Chow has had some success in applying information theoretic methods to this type of problem. Perhaps other methods could be fruitful as well.

BIBLIOGRAPHY

BIBLIOGRAPHY

[A-1] Abramson, N., and D. Braverman, "Learning to Recognize Patterns in a Random Environment," IRE Trans., Vol. IT-8, pp. 558-563, 1962.

[A-2] Aizerman, M. A., E. M. Brauerman, and L. I. Rozonoer, "Theoretical Foundations of the Potential Function Method in Pattern Recognition Learning," Automatica i Telemekhanika, Vol. 25, June 1964, trans. Jan. 1965, pp. 821-837.

[A-3] Albert, A. "A Mathematical Theory of Pattern Recognition," Ann. Math. Stat., Vol. 34, pp. 284-299, March, 1963.

[A-4] Albert, Arthur E., and Leland A. Gardner, Jr. Stochastic Approximation and Nonlinear Regression, MIT Press, Cambridge, 1967.

[A-5] Amari, Shunichi, "A Theory of Adaptive Pattern Classifiers," IEEE Trans., Vol. EC-16, No. 3, pp. 299-307, June 1967.

[A-6] Anderson, T. W., and R. R. Bahadur, "Classification Into Two Multivariate Normal Distributions with Different Covariance Matrices," Ann. Math. Stat., Vol. 33, pp. 420-431, June 1962.

[A-7] Aoki, Masanao, Optimization of Stochastic Systems, Academic Press, New York, 1967.

[B-1] Bashkirov, O. A., E. M. Braverman, and I. B. Muchnik, "Potential Function Algorithms for Pattern Recognition Learning Machines," Automatic Machines and Remote Control, Vol. 25, No. 5, pp. 629-631, March, 1964, trans. from Aut. i Telemekh., Vol. 25, No. 5, pp. 692-695, May 1964.

[B-2] Blaydon, Colin C.,"Recursive Algorithms for Pattern Classification," Thesis, Harvard University, Div. of Engineering and Applied Physics, March 1967.

[B-3] Bledsoe, W. W., "Some Results on Multicategory Pattern Recognition," J. ACM, Vol. 13, No. 2, pp. 304-316, April 1966.

[C-1] Chadwick, Henry D., and Ludwik Kurz, "Two Sequential Nonparametric Detection Procedures," Info. & Control, Vol. 13, No. 5, pp. 403-428, November 1968.

[C-2] Chien, Y. T., and K. S. Fu, "On Bayesian Learning and Stochastic Approximation," IEEE Trans., Vol. SSC-3, No. 1, pp. 28-38, June 1967.

[C-3] Chien, Y. T., and K. S. Fu, "Selection and Ordering of Feature Observation in a Pattern-Recognition System," Info. & Control, Vol. 12, Nos. 5/6, pp. 394-414, May-June 1968.

[C-4] Choi, Keewhan, "Estimators for the Parameters of a Finite Mixture of Distributions," Annals of the Inst. of Stat. Math., Tokyo, Vol. 21, No. 1, pp. 107-116, 1969.

[C-5] Choi, Keewhan, "Empirical Bayes Procedure for (Pattern) Classification with Stochastic Learning," Ann. of the Inst. of Stat. Math., Tokyo, Vol. 21, No. 1, pp. 117-125, 1969.

[C-6] Chow, C. K., "A Recognition Method Using Neighbor Dependence," IRE Trans., Vol. EC-11, pp. 683-690, Oct. 1962.

[C-7] Chow, C. K., "Statistical Independence and Threshold Functions," IEEE Trans., Vol. EC-14, No. 1, pp. 66-68, Feb. 1965.

[C-8] Chow, C. K., and C. N. Kiu, Approximating Discrete Probability Distributions with Dependence Tree, IBM Research Report RC-1816, May 4, 1967.

[C-9] Chu, J. T., "Optimal Decision Functions for Computer Character Recognition," J. of ACM, Vol. 12, No. 2, pp. 213-226, 1965.

[C-10] Chu, J. T., and J. C. Chueh, "Error Probability in Decision Functions for Character Recognition," J. of ACM, Vol. 14, No. 2, pp. 273-280, April 1967.

[C-11] Cooper, P. W., "The Hyperplane in Pattern Recognition," Cybernetica, Vol. 4, pp. 215-218, 1962.

[C-12] Cooper, D. B., and P. W. Cooper, "Nonsupervised Adaptive Signal Detection and Pattern Recognition," Info. & Control, Vol. 7, pp. 416-444, 1964.

[C-13] Cooper, P. W., "Hyperplanes, Hyperspheres, and Hyperquadrics as Decision Boundaries," Computer and Information Sciences, Tou and Wilcox, Eds., Spartan, Washington, D. C., 1964.

[C-14] Cover, Thomas M., "Geometrical and Statistical Properties of Systems of Linear Inequalities with Applications in Pattern Recognition," IEEE Trans., Vol. EC-14, No. 3, pp. 326-334, June 1965.

[D-1] Deely, J. J., and R. L. Kruse, "Construction of Sequences Estimating the Mixing Distribution," Ann. Math. Statist., Vol. 39, No. 1, pp. 286-288, 1968.

[D-2] Dubes, R. C., The Theory of Applied Probability, Prentice-Hall, Englewood Cliffs, N. J., 1968.

[D-3] Dubes, R. C., and P. J. Donoghue, "Bayesian Learning in Markov Chains with Observable States," Interim Report No. 5, Contract No. AFOSR-1023-67B, Division of Engineering Research, Michigan State University.

[D-4] Dubes, R. C., and E. Panayirci, "Pattern Recognition with Continuous Parameter, Observable Markov Chains," Interim Report No. 10, Contract No. AFOSR-1023-67B, Division of Engineering Research, Michigan State University.

[D-5] Duda, R. O., and H. Fossum, "Pattern Classification by Iteratively Determined Linear and Piecewise Linear Discriminant Functions," IEEE Trans., Vol. EC-15, No. 2, pp. 220-232.

[F-1] Farrell, J. L., and J. C. Murtha, "Statistical Band Synchronization in Digital Communication," Proceedings of the Univ. of Missouri, Rolla, Mervin J. Kelly Communications Conference, Oct. 1970, pp. 3-5-1 through 3-5-6.

[F-2] Feller, W., An Introduction to Probability Theory and Its Applications, John Wiley & Sons, Inc., New York, 1957.

[F-3] Friedman, H. D., "On the Expected Error in the Probability of Misclassification," Proc. IEEE, Vol. 53, pp. 658-659, June 1965.

[F-4] Fraser, D. A. S., Non-Parametric Methods in Statistics, Wiley, New York, 1957.

[F-5] Fu, K. S., Sequential Methods in Pattern Recognition and Machine Learning, Academic Press, New York, 1968.

[G-1] Good, I. J., The Estimation of Probabilities, Research Monograph No. 30, MIT Press, Cambridge, 1965.

[H-1] Hancock, J. C., and T. L. Stewart, "Parameter Estimation with Unknown Symbol Synchronization," Thesis, Purdue University, January 1967.

[H-2] Hancock, J. C., and P. A. Wintz, Signal Detection Theory, McGraw-Hill, New York, 1966.

[H-3] Harnett, W. E., "Generalization of Tests for Certain Properties of Variable-Length Codes," Info. & Control, Vol. 13, No. 1, pp. 20-24, July 1968.

[H-4] Hellman, M. E., and T. M. Cover, "Learning with Finite Memory, Ann. Math. Stat., Vol. 41, No. 3, pp. 765-782, June 1970.

[H-5] Hellman, Martin E., and Josef Raviv, "Probability of Error, Equivocation, and the Chernoff Bound," IEEE Trans., Vol. IT-16, No. 4, pp. 368-372, July 1970.

[H-6] Hilborn, C. G., and D. G. Lainiotis, "Optimum Unsupervised Learning Multicategory Dependent Hypothesis Pattern Recognition," IEEE Trans., Vol. IT-14, No. 3, pp. 468-470, May 1968.

[H-7] Ho, and Agrawala, "Summary of the State of the Art in Pattern Recognition," Proc. IEEE., Vol. 56, No. 12, pp. 2101-2114, Dec. 1968.

[H-8] Hoeffding, W., and H. Robbins, "The Central Limit Theorem for Dependent Random Variables, Duke Math J., Vol. 15, pp. 773-780, 1948.

[I-1] Irani, K. B., "A Finite-Memory Adaptive Pattern Recognizer," IEEE Trans., Vol. SSC-4, No. 1, pp. 2-11, March 1968.

[K-1] Kadota, T. T., and L. A. Shepp, "On the Best Set of Linear Observables for Discriminating Two Gaussian Signals," IEEE Trans., Vol. IT-13, No. 3, pp. 278-284, April 1967.

[K-2] Kailath, Thomas, "The Divergence and Bhattacharyya Distance Measures in Signal Selection," IEEE Trans., Vol. COM-15, No. 1, pp. 52-60, February 1967.

[K-3] Kazmierczak, H., and K. Steinbuch, "Adaptive Systems in Pattern Recognition," IEEE Trans., Vol. EC-12, No. 6, pp. 822-835, Dec. 1963.

[K-4] Keehn, Daniel G., "A Note on Learning for Gaussian Properties," IEEE Trans., Vol. IT-11, No. 1, pp. 126-132, Jan. 1965.

[K-5] Kobayashi, H., and J. B. Thomas, "Distance Measures and Related Criteria," Proceedings of the Fifth Allerton Conference on Circuit and System Theory, pp. 491-500, Oct. 1967.

[K-6] Korsh, James F. "On Decisions and Information Concerning an Unknown Parameter," Info. & Control, Vol. 16, pp. 123-127, 1970.

[K-7] Ku, Harry S., and Solomon Kullback, "Approximating Discrete Probability Distributions," IEEE Trans., Vol. IT-15, No. 4, pp. 444-447, July 1969.

[L-1] Lainiotis, D. G., "Optimal Feature Extraction in Pattern Recognition," IEEE Intern. Info. Theory Symp. Abstracts, San Remo, Italy, Sept. 1967.

[L-2] Lainiotis, Demetrios G., "On a General Relationship Between Estimation, Detection, and the Bhattacharyya Coefficient," IEEE Trans., Vol. IT-15, No. 4, pp. 504-505, July 1969.

[L-3] Lainiotis, D. G. "A Class of Upper Bounds on Probability of Error for Multihypothesis Pattern Recognition," IEEE Trans., Vol. IT-15, No. 6, pp. 730-731, Nov. 1969.

[L-4] Lainiotis, D. G., and S. K. Park, "Probability of Error Bounds," IEEE Trans., Vol. SMC-1, No. 2, pp. 175-178, April 1971.

[M-1] Martin, J. J., Bayesian Decision Problems and Markov Chains, Wiley, New York, 1967.

[M-2] Martin, James. Telecommunications and the Computer, Prentice-Hall, Englewood Cliffs, N.J., 1969.

[M-3] McBride, Alan L., and Andrew P. Sage, "Optimum Estimation of Bit Synchronization," IEEE Trans., Vol. AES-5, No. 3, pp. 525-536, May 1969.

[M-4] McBride, Alan L., and Andrew P. Sage, "On Discrete Sequential Estimation of Bit Synchronization," IEEE Com., Vol. 18, No. 1, Feb. 1970.

[M-5] McLendon, J. R. "A Pseudo Bayes Approach to Digital Detection and Likelihood Ratio Computation," Thesis, Southern Methodist University, Dec. 1969.

[M-6] Meisel, William S. "Potential Functions in Mathematical Pattern Recognition," IEEE Trans., Vol. C-18, No. 10, pp. 911-918, Oct. 1969.

[M-7] Middleton, D., "On the Detection of Random Signals in Additive Normal Noise--Part 1," IREE Trans., Vol. IT-3, No. 2, pp. 86-122, June 1957.

[M-8] Middleton, D., An Introduction to Statistical Communication Theory, McGraw-Hill, New York, 1960.

[N-1] Nagy, G., "State of the Art in Pattern Recognition," Proc. IEEE., Vol. 56, pp. 836-862, 1968.

[N-2] Nilsson, N. J., Learning Machines, McGraw-Hill, New York, 1965.

[P-1] Panayirci, E., "Bayesian Decision Making and Learning for Continuous-Time Markov Systems," Ph.D. dissertation, Michigan State University, 1970.

[P-2] Papoulis, A., Probability, Random Variables, and Stochastic Processes, McGraw-Hill, New York, 1965.

[P-3] Patrick, E. A., and J. C. Hancock, "Nonsupervised Sequential Classification and Recognition of Patterns," IEEE Trans., Vol. IT-12, No. 3, pp. 362-372, July 1966.

[P-4] Pitt, J. M., and B. F. Womack, "Additional Features of an Adaptive, Multicategory Pattern Classification System," IEEE Trans., Vol. SSC-5, No. 3, pp. 183-191, July 1969.

[R-1] Raviv, J., "Decision Making in Incompletely Known Stochastic Systems," Int. J. Eng. Sci., Vol. 3, pp. 119-140, 1965.

[R-2] Raviv, J. "Decision Making in Markov Chains Applied to the Problem of Pattern Recognition," IEEE Trans., Vol. IT-3, No. 4, pp. 536-551, Oct. 1967.

[R-3] Rényi, Alfréd, "On the Amount of Information Concerning an Unknown Parameter in a Sequence of Observations," Magyar Tud. Akad. Mat. Kutató Int. Közl., Vol. 9, pp. 617-625, 1964.

[R-4] Rényi, A., "On the Amount of Missing Information and the Neyman-Pearson Lemma," Research Papers in Statistics Festschrift for J. Neyman, Wiley, New York, 1966.

[R-5] Rényi, A. "On Some Basic Problems of Statistics from the Point of View of Information Theory," Proc. 5th Berkeley Symp. on Math. Stat. and Prob., Vol. 1, Berkeley, Calif., U. of Cal. Press, 1967.

[R-6] Robbins, Herbert, "The Empirical Bayes Approach to Statistical Decision Problems," Ann. Math. Statist., Vol. 35, pp. 1-20, 1964.

[R-7] Rolph, John E., "Bayesian Estimation of Mixing Distribution," Columbia Univ.

[S-1] Sebestyen, George S., Decision-Making Processes in Pattern Recognition, Macmillan, New York, 1962.

[S-2] Sebestyen, G., and J. Edie, "An Algorithm for Non-Parametric Pattern Recognition," IEEE Trans., Vol. EC-15, No. 6, pp. 908-915, Dec. 1966.

[S-3] Signori, D. T., Jr., "Estimation and Adaptive Decision Making for Partially Observable Markov Systems," Ph.D. dissertation, Michigan State University, 1968.

[S-4] Specht, Donald F., "Generation of Polynomial Discriminant Functions for Pattern Recognition," IEEE Trans., Vol. EC-16, No. 3, June 1967.

[S-5] Spragins, J. D. "Learning Without a Teacher," IEEE Trans., Vol. IT-12, No. 2, pp. 273-230, April 1966.

[S-6] Spragins, J. D. Reproducing Distributions for Machine Learning, Stanford Electronic Lab, Tech. Report No. 6103-7, Nov. 1963.

[T-1] Tallis, G. M., "The Identifiability of Mixtures of Distributions," J. Appl. Prob., Vol. 6, pp. 389-398, 1969.

[T-2] Teicher, H., "Identifiability of Mixtures," Ann. Math. Stat., Vol. 32, pp. 244-248, 1961.

[T-3] Toussaint, G. T., "Some Upper Bounds on Error Probability for Multiclass Feature Selection," IEEE Trans., Vol. C-20, No. 8, pp. 943-944, Aug. 1971.

[T-4] Tucker, Howard G., A Graduate Course in Probability, Academic Press, New York, 1967.

[W-1] Wainstein, L. A., and V. D. Zubakov, Extraction of Signals from Noise, Prentice-Hall, Englewood Cliffs, N.J., 1962.

[Y-1] Yakowitz, Sidney J., "A Consistent Estimator for the Identification of Finite Mixtures," Ann. Math. Stat., Vol. 40, No. 5, pp. 1728-1735, 1969.

[Y-2] Yakowitz, Sidney J., "Unsupervised Learning and the Identification of Finite Mixtures," IEEE Trans., Vol. IT-16, No. 3, pp. 330-338, May 1970.

[Y-3] Yakowitz, Sidney J., and John D. Spragins, "On the Identifiability of Finite Mixtures," Ann. Math. Stat., Vol. 39, No. 1, pp. 209-214, 1968.

[Y-4] Yau, S. S., and J. M. Schumpert, "Design of Pattern Classifiers with the Updating Property Using Stochastic Approximation Techniques., IEEE Trans., Vol. C-17, No. 9, pp. 861-872. Sept. 1968.