ESTIMATION AND ADAPTIVE DECISION MAKING FOR PARTIALLY OBSERVABLE MARKOV SYSTEMS

Thesis for the Degree of Ph. D.
MICHIGAN STATE UNIVERSITY
DAVID T. SIGNORI, JR.
1968

This is to certify that the

thesis entitled

ESTIMATION AND ADAPTIVE DECISION MAKING
FOR PARTIALLY OBSERVABLE MARKOV SYSTEMS

presented by

David T. Signori, Jr.

has been accepted towards fulfillment
of the requirements for

Ph.D. degree in E.E.

Major professor

Date Sept. 25, 1968

O-169

ABSTRACT

A Partially Observable Markov System (POMS) is a discrete state, discrete time system whose state activity is described by a Markov chain. The states of the system cannot be observed directly, but "noisy" observations are available.

The main problem considered is that of determining rules for making decisions about system states when the conditional densities of observed random variables given the state of the system are characterized by a set of unknown parameters. Furthermore, it is desired that, as more observations are taken, these rules converge to the rule that would be used if the parameters were known.

An iterative, optimal (minimum Bayes risk), decision rule is derived for making decisions concerning the state of the system at a given time on the basis of available observations. This rule has the capability of using future observations as well as past observations. An optimal rule is also established for determining to which class the state of the system belongs among a set of non-communicating classes of states and an optimal, adaptive estimator is constructed for the parameters associated with the active class. Conditions are established under which these rules perform effectively.

A variety of consistent estimators are constructed for the unknown parameters, yielding a class of suboptimum rules. The basic estimation problem is a nonsupervisory one involving the resolution of mixtures. However, unlike previous work, the observation process is dependent and nonstationary. A general strategy is established

for extending estimation techniques developed for the case of independent identically distributed observations to this problem. The results apply also to the non-parametric case and the case with unknown transition matrix.

The model under study here corresponds directly to that of a Pattern Recognition System with Markov dependent pattern activity. However, several communication systems of interest can be shown to be POMS. These include systems with a Markov encoder, intersymbol interference, unknown synchronization, signals with random arrival times, and combinations of the foregoing. In all of these systems, the observations are dependent and the design of adaptive detectors is generally difficult. However, by formulating the problem as one of decision making for a POMS, optimal and suboptimal detectors as well as conditions for effective operation follow easily and in a unified manner.

ESTIMATION AND ADAPTIVE DECISION MAKING FOR PARTIALLY OBSERVABLE MARKOV SYSTEMS

By

David T. Signori, Jr.

A THESIS

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

DOCTOR OF PHILOSOPHY

Department of Electrical Engineering

1968

To

My Mother and Father

ACKNOWLEDGEMENTS

It is with great pleasure that the author thanks his thesis advisor Dr. R.C. Dubes for his guidance and encouragement during the course of this research

Thanks are also due to Dr. H. Salehi and Dr. R.B. Zemach for their suggestions concerning some mathematical questions and to Dr. H.E. Koenig and Dr. G.L. Park for their interest in this work.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

Introduction

A fundamental requirement of classical engineering techniques for the design of control and information processing systems is the existence of a fully specified model of the system under study including all the factors that influence system performance. However, factors such as unpredictable changes in environment, drift in system parameter values due to component aging, or difficulties in measuring relevant quantities often make the development of an accurate model impractical. This problem has motivated the study of adaptive decision making devices such as controllers and detectors which can achieve a desired goal despite some degree of ignorance concerning the underlying model. Such devices are characterized by the fact that they improve their performance on the basis of past experience. In effect they "learn" the additional information needed to complete the model in terms of which their task is defined.

Workable adaptive schemes have been proposed for several types of communication and control systems [A-1][H-4][M-2][S-4]. In particular, a large class of problems that arise from analyzing such systems can be posed as problems in Mathematical Statistics. In such problems Decision Theory provides the mathematical form of the adaptive device and a mathematical criterion for evaluating its performance. The learning process is related to the well-defined problem of estimating the unknowns in a partially specified statistical structure. The only prerequisite for using the powerful tools of Mathematical Statistics is

the existence of a meaningful statistical model.

The basic model under study in this Thesis is that of a discrete state, discrete time system whose state activity is described by a Markov chain. The system states cannot be observed directly because of an imperfect observation mechanism that can be accounted for statistically. This system will be referred to as a Partially Observable Markov System (POMS). Such systems arise frequently in Pattern Recognition, Signal Detection, and Operations Research [D-3][K-1][R-3]. The model for a POMS is precisely defined in Sec. 1.1.

The main decision problem associated with a POMS is one of establishing rules for taking effective action concerning the states of the system on the basis of available observations. In Sec. 1.2 optimal decision making is defined for a POMS when all quantities in its model are assumed known. The structure of the resulting rule is discussed along with its computational feasibility, a basic consideration throughout this study.

This Thesis deals with various aspects of decision making for a POMS when the model is not completely specified (not all the quantities in the model are known). Of primary interest is the problem of extracting from the observations information concerning the model unknowns. Hopefully, such information can be used to construct adaptive decision rules or rules which perform almost as well (in some well defined manner) as the optimal rules of Sec. 1.2 where the model is completely known. The problem of learning the unknowns in a POMS is discussed in Sec. 1.3. Previous work related to this problem is listed in Sec. 1.4 and in Sec. 1.5 the Thesis objectives are explicitly stated.

1.1 PARTIALLY OBSERVABLE MARKOV SYSTEMS

The basic model considered in the Thesis is established in this section. The model is composed of two random processes. The first is a discrete-time, finite-dimensional Markov chain which cannot be observed. The second is an observable process with the property that the random variable describing the observations at a given time has a distribution which depends on the state of the chain at that time. The model corresponds to that of a system whose states cannot be observed but must be monitored indirectly through a "noisy" observation mechanism, which suggests the name Partially Observable Markov System (POMS).[1]

More specifically, the state activity of the system is described by a first order homogeneous Markov chain; that is, a sequence of random variables $\{\lambda_N;\ N = 1,2,\ldots\}$ taking values in a finite state space $\Lambda = \{1,2,\ldots,M\};\ M < \infty$ and satisfying the Markov Property. Namely, if $P(\cdot)$ is a probability measure defined on the same sample space as the sequence $\{\lambda_N\}_1^\infty$ then

$$P(\lambda_N = j \,|\, \lambda_{N-1} = i,\ldots,\lambda_1 = k) = P(\lambda_N = j \,|\, \lambda_{N-1} = i) = p_{ij}$$

$$\forall N > 1 \quad \text{and} \quad i,j = 1\ldots M \qquad (1.1.1)$$

where p_{ij} is the probability the system is in state j at time N given it was in state i at time $N-1$. Hence, knowledge of the last state summarizes the past history of the system. The probability state vector of the system is defined by

$$\underline{p}_N = [P(\lambda_N=1),\ldots,P(\lambda_N=M)] \qquad (1.1.2)$$

---

[1] This is a generalization to a continuous observation space of what has been previously referred to as a POMS [D-3],[K-1].

where $P(\lambda_N=i)$ is the probability that the system is in state i at time N. Then

$$\underline{p}_N = \underline{p}_1 P^{N-1} = \underline{p}_{N-1} P \tag{1.1.3}$$

where $P = [p_{ij}]$ is a stationary transition matrix and $\underline{p}_1$ is an initial probability state vector at time 1. Then $\underline{p}_1$ and P are sufficient to summarize the prior knowledge (knowledge before any observations are taken) of the state activity of the system.

When the past history of the system is summarized by knowledge of the last k states, the describing random process is termed a k^{th} order Markov chain. Since any k^{th} order chain can be reduced to a 1^{st} order chain, the above chain implies higher order chains [D-2].

The observation process is defined by a sequence of random variables $\{X_N\}_1^\infty$; X_N, the random variable observed at time N, takes values in a finite-dimensional Euclidean space and has a density function $f_i(\cdot)$ when it is known the system is in state i at time N. That is, $f_i(\cdot)$ is the conditional density of the observations given the system is in state i. Since the states of the system are unknown, X_N has the global density

$$p_N(X) = \sum_{i=1}^{M} f_i(X)\ P(\lambda_N=i) \tag{1.1.4}$$

which is referred to as a finite mixture with component densities $\{f_i(\cdot)\}_1^M$ and mixing parameters $\underline{p}_N$ [T-1]. The sequence $\{X_N\}_1^\infty$ is assumed state conditionally independent. This implies, for example,

$$p(X_N, X_{N+1} | \lambda_N=i, \lambda_{N+1}=j) = f_i(X_N) f_j(X_{N+1}) \tag{1.1.5}$$[2]

---

[2] $p(\cdot)$ will denote probability density with X_i indicating both the random variables and the value it takes on.

and hence the joint density of any number of random variables in the observation process can be constructed from the component densities and the system probabilities p_1 and P.

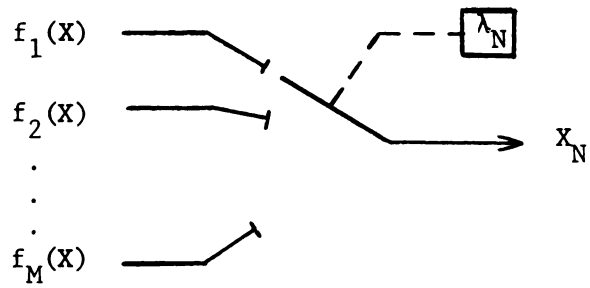

FIG. 1.1.1 A Schematic of a Partially Observable Markov System

The entire model is illustrated by the sampling scheme in Fig. 1.1.1. At time N a sample is taken according to a density determined by the state of the system at time N.

1.2 OPTIMAL DECISION MAKING FOR A POMS

When the model is completely specified, Decision Theory provides a means of generating optimal strategies for action relative to the states of the system. In this section the type of decision rules of interest in the Thesis are illustrated along with important properties of this class of rules.

In order to define the decision problem, the following elements of Decision Theory are introduced. The action space or set of allowable actions that can be taken at a given time is denoted by $A = \{a_1,\ldots,a_r\}$, $r < \infty$ with generic element a; $L(\cdot,\cdot)$ is a non-negative loss function defined on $A \times \Lambda$ with $L(a,i)$ denoting the loss incurred when action

a is taken and the system is in state i; $X^m = [X_1,\ldots,X_m]$ represents the set of observations obtained up to time m. The problem, then, is to find a nonrandomized decision function δ^N mapping X^m to A such that the risk

$$R(\delta^N) = E\, L(\delta^N(X^m),\lambda_N) \tag{1.2.1}$$

is a minimum. The expectation is taken with respect to X^m and λ_N.

If, for example, r = M and the action a_i is "say λ_N=i" and $L(i,j) = 1 - \Delta_{ij}$ (the 0-1 loss function)[3] the optimal decision rule, δ^N_*, is given by the Bayes decision rule

$$\delta^N_*(X^m) = i \quad \text{if} \quad P(\lambda_N=i/X^m) \geq P(\lambda_N=j/X^m) \quad \forall j\neq i \tag{1.2.2}$$

where $P(\lambda_N=i/X^m)$ is the conditional probability that the system is in state i at time N given observation X^m [R-3]. Using the assumption of state conditional independence the above rule can be written iteratively. For example, if m = N it follows from Bayes rule and the Markov property that

$$P(\lambda_N=i/X^N) = \frac{f_i(X_N)P(\lambda_N=i/X^{N-1})}{p(X_N/X^{N-1})} \tag{1.2.3}$$

where

$$P(\lambda_N=i/X^{N-1}) = \sum_1^M p_{ki}P(\lambda_{N-1}=k/X^{N-1}) \tag{1.2.4}$$

and

$$P(X_N/X^{N-1}) = \sum_1^M f_i(X_N)P(\lambda_N=i/X^{N-1}). \tag{1.2.5}$$

---

[3] Δ_{ij} denotes the Kronecker delta.

The iterative scheme operates in two basic steps. The state posterior probability $P(\lambda_{N-1}=i/X^{N-1})$ computed at step N-1 is projected from the transition probabilities in (1.2.4) to $P(\lambda_N=i/X^{N-1})$ which serves as a prior probability for the state activity at time N before the N^{th} observation is available. This in turn is converted in (1.2.3) to a posterior probability using X_N and the Bayes Rule.

It is worth noting that, because of the Markov dependencies between the states of the system at different times, observations are in general dependent. Consequently, observations at one time may contain information about the states at other times. This point is reflected in the above decision rule by the fact that both past and present observations are used to make decisions about the state of the system at a given time. By the same token the sequence of decision made by $\{\delta_*^i\}_1^N$ can usually be improved if X^N is used to classify all past states simultaneously. However, rules of this type lead to memory requirements which grow linearly and exponentially with the length of the observed sequence and decisions are not available for immediate use [C-1].

The type of decision rules established in this section can be thought of as a class of on-line rules with fixed memory but changing structure. The case where $m > N$, termed a look ahead mode, is an attempt to improve performance by increasing memory by a fixed amount.

1.3 UNSUPERVISED LEARNING

If the model for a POMS is only partially specified the previously established decision rules can be considered functions of the

unknowns in the model. For example, the transition matrix and/or the component densities may be unknown or may contain unknown parameters. To make effective decisions under such circumstances information about these unknowns must be extracted from the observations. Since the states of the system are unobservable the unknowns appear in a mixture. The process of estimating or approximating these unknowns is commonly referred to as unsupervised learning. This is in contrast to the case in which, by some external means, the states of the system are known for a fixed period of time and each of the component densities and the transition matrix can be determined separately or with supervision [B-1] [M-1][H-4].

As an example of the above class of estimation problems a POMS is considered with a transition matrix which has identical row vectors. That is

$$p_{ij} = q_j \quad j=1,2,\ldots,M \tag{1.3.1}$$

If $\underline{p}_1$ is given by a row of P, the probability state vector $\underline{p}_N$ is independent of time and X_N, the random variable observed at time N, has density

$$p(X) = \sum_1^M q_j f_j(X) \quad \forall_N \tag{1.3.2}$$

Then the state conditional independence assumption implies $\{X_i\}_1^\infty$ is a sequence of independent identically distributed (i.i.d.) random variables. This is a well-studied case and the tools of classical estimation theory have been used to develop several mixture resolving techniques for various degrees of uncertainty about (1.3.2) [S-6][H-2].

Some of these methods are discussed in Chapters II and III.

However, for a general transition matrix and initial probability vector the observations are neither independent nor identically distributed and the above techniques do not apply directly.

1.4 REVIEW OF THE LITERATURE

Research related to adaptive decision making for a POMS has been motivated by the need to handle decision problems in which the usual independence and stationarity assumptions mentioned in Sec. 1.3 do not hold. The results outlined in this section can be categorized according to what is assumed to be unknown in the model.

For the case in which the only unknowns are parameters in the component densities, most of the work consists of attempts to design optimal (Minimum Bayes Risk) and suboptimal adaptive detectors for communication systems wherein the signal is unknown but, for various reasons, the observations are dependent even when the signal is known.[4] Some examples of conditions which result in dependent observations are intersymbol interference or signal overlap due to channel memory (Chang [H-1]), unknown symbol synchronization between the transmitter and receiver (Stewart [H-3]) and random signal arrival times (Stewart [H-3] and Nolte [N-1]).

The main problem associated with these examples is to establish conditions under which the unknown parameters can be learned. In the first two examples Chang and Stewart were able to establish such conditions because they assumed signal activity at the transmitter was independent from one time interval to the next.[5] Their results do not

---

[4] Optimal and suboptimal adaptive decision rules are defined at the beginning of Chapters II & III respectively. The correspondence between a POMS and a communication system is made in Chapter IV.

[5] This corresponds to assuming a special form for the transition

apply when such dependencies exist. In the example of signals with random arrival times Stewart was unable to prove convergence of his estimates and Nolte does not treat the problem. These examples are discussed in more detail in Chapter IV.

For the case in which the component densities are assumed known but the transition matrix P is unknown Raviv [R-2] constructed a class of adaptive decision rules using an estimate of P and only part of the past observations. He established conditions under which P can be estimated and developed some properties of the observation process for a large class of POMS. These properties are stated in Chapter III.

Recently, Patrick [P-2] and Hilborn and Lainiotus [H-5] have made some general observations concerning non-supervisory problems with non-stationary, dependent observations. The work in this Thesis related to their results was done independently and deals with a particular model which yields more specific results.

1.5 THESIS OBJECTIVES

For most of the work in this Thesis it is assumed that the transition matrix P is known and that the component densities are known to within a parameter.[6] For this case, the Thesis objectives can be stated generally as follows:

1. To find a class of optimal adaptive decision rules with properties similar to those of the rules of Sec. 1.2.
2. To show that, under appropriate assumptions, virtually all the mixture resolving techniques developed for the i.i.d.

---

[6] The estimation of P and $f_i(X)$ for the nonparametric case is discussed briefly in Chapter III.

case defined in Sec. 1.3 can be extended to the dependent, non-stationary case considered here.

3. To show that a variety of decision problems arising in communication systems, including those of Sec. 1.4, can be easily solved by considering them as decision making problems for a POMS.

Fulfillment of Objective 1 provides a reference for evaluation of any related adaptive scheme. Objective 2 implies conditions under which the unknown parameters can be learned and leads to a class of suboptimum decision rules. Objective 3 suggests that the model of a POMS is a very versatile one, providing a unifying approach to a class of communications problems.

Objectives 1, 2, and 3 are pursued in Chapters II, III, and IV respectively. In Chapter V the main results of the Thesis are outlined and problems that need further study are discussed.

CHAPTER II

OPTIMAL ADAPTIVE DECISION MAKING

When the component densities of a Partially Observable Markov System (POMS) are specified to within a parameter set and a prior distribution[7] summarizing initial knowledge about the unknown parameters is available then Bayes decision-theoretic techniques can generally be used to establish decision rules which are optimal against prior information and a given cost function. Furthermore, under appropriate conditions, the fixed but unknown value of the parameter is learned from the observations and the decision rule adapts or converges to what the optimal rule would be if the true parameter value were known. Patrick [H-2] derived the optimal decision rule for the i.i.d. case.

In Sec. 2.1, the decision problem is defined and in Sec. 2.2 the corresponding optimal decision rule is derived. In Sec. 2.3, the basic components of the optimal rule are generated recursively and in Sec. 2.4 the structure and computational feasibility of the iterative scheme are discussed. In Sec. 2.5, the learning properties of the rule are discussed and in Sec. 2.6 some inference problems related to that defined in Sec. 2.1 are treated. Finally, in Sec. 2.7, the main results of the chapter are summarized.

2.1 THE DECISION PROBLEM

In this section, the decision problem under consideration in this chapter is defined. The basic elements of the problem are a POMS with an unknown parameter set and a prior density for the parameters,

---

[7] The standard Bayesian technique of treating the unknown parameter as a random variable will be employed.

a class of decision functions and a criterion for evaluating the performance of these functions.

As in Sec. 1.1, the state activity of the POMS is described by a Markov chain $\{\lambda_N\}_1^\infty$ with transition matrix P and initial probability state vector $\underline{p}_1$. The component densities are characterized by a parameter vector in the following manner. When the system is known to be in state i and the parameter B_i is given, X_N, the random variable observed at time N, has density $f(\cdot/B_i)$.[8] The random vector $B = [B_1 \ldots B_M]$ takes on an unknown value B_0 according to the prior density $p_0(B)$ and maintains this value throughout system operation. The basic assumptions are as follows.

1. The observations $X^m = [X_1 \ldots X_m]$ are state-parameter-conditional independent. This implies

$$p(X_m \ldots X_1/\lambda_m = i, \ldots \lambda_1 = j, B) = p(X_m/\lambda_m = i, B_i) \ldots p(X_1/\lambda_1 = j, B_j)$$
$$= f(X_m/B_i) \ldots f(X_1/B_j) \qquad (2.1.1)$$

or

$$p(X_m/\lambda_m = i, X^{m-1}, B) = p(X_m/\lambda_m = i, B_i) = f(X_m/B_i) \qquad (2.1.2)$$

2. For each N, the random variables B and λ_N are independent. That is, the parameter values do not affect system state activity.

If, as in Sec. 1.2, A is the action space, the class of allowable decision rule, D, is the set of all non-randomized functions mapping the space of observations X^m to A. If $\delta^N \in D$ denotes a decision rule for taking action relative to the state of the system at

---

[8] The random variables X_i and B_i take values in a finite-dimensional Euclidean vector space.

time N and $L(\cdot,\cdot)$ is a nonnegative loss function, the corresponding risk is given by

$$R(\delta^N) = E\, L(\delta^N(X^m), \lambda_N) \tag{2.1.3}$$

where the expectation is taken with respect to X^m, λ_N, and B.

The problem is to find the optimal rule, δ^N_*, which is defined by

$$R(\delta^N_*) \leq R(\delta^N) \quad \forall\, \delta^N \in D \tag{2.1.4}$$

That is, the minimum risk decision rule for taking action concerning the state of the system at time N on the basis of m observations is to be found.

2.2 DERIVATION OF THE DECISION RULE

In this section, the optimum decision rule is derived for the problem defined in Sec. 2.1. The development involves the use of conditional risks which emphasize the role of prior information in constructing the total risk.

With the decision rule $\delta \in D$, given observation X^m, and given parameter B the average loss is

$$R[\delta(X^m)/X^m, B] = \sum_{i=1}^{M} L[\delta(X^m), i]\, P(\lambda_N = i/X^m, B) \tag{2.2.1}$$[9]

Equation (2.2.1) is referred to as the sample-parameter-conditional risk. The parameter-conditional risk is given by

$$R(\delta/B) = \int R[\delta(X^m)/X^m, B]\, p(X^m/B)\, dX^m \tag{2.2.2}$$

---

[9] The superscript in δ^N has been dropped for convenience.

and the total risk, (2.1.3), is

$$R(\delta) = \int R(\delta/B)\, p_0(B)\, dB \tag{2.2.3}$$

Substituting (2.2.1) and (2.2.2) into (2.2.3) and interchanging the order of integration yields

$$R(\delta) = \int R[\delta(X^m)/X^m]\; p(X^m)\, dX^m \tag{2.2.4}$$

where

$$R[\delta(X^m)/X^m] = \sum_{i=1}^{m} L[\delta(X^m), i] P(\lambda_N = i/X^m) \tag{2.2.5}$$

and

$$P(\lambda_N = i/X^m) = \int P(\lambda_N = i/X^m, B)\, p(B/X^m)\, dB \tag{2.2.6}$$

$$p(B/X^m) = p(X^m/B)\, p_0(B)/p(X^m) \tag{2.2.7}$$

$$p(X^m) = \int p(X^m/B)\, p_0(B)\, dB \tag{2.2.8}$$

Since $\delta(X^m) = a$ for some $a \in A$, and $L(\cdot,\cdot)$ and $p(X^m)$ are non-negative it follows that

$$R(\delta_*) = \inf_{\delta \in D} R(\delta) = E\, R[\delta_*(X^m)/X^m] \tag{2.2.9}$$[10]

where

$$\delta_*(X^m) = a_i \quad \text{if} \quad R(a_i/X^m) \leq R(a_j/X^m) \quad \forall\, j \neq i \tag{2.2.10}$$

In particular, when $r = M$, a_i is the action "say $\lambda_N = i$" and the loss function is given by $L(a_i, j) = 1 - \Delta_{ij}$ (0-1 loss function) then the decision rule becomes a minimum probability of error rule δ_e defined by

$$\delta_e(X^m) = i \quad \text{if} \quad P(\lambda_N = i/X^m) \geq P(\lambda_N = j/X^m) \quad \forall\, i \neq j \tag{2.2.11}$$

---

[10] Minimum sample conditional risk implies minimum total risk.

The basic elements of the above decision rules are the state posterior probabilities given by (2.2.6). These probabilities are obtained by averaging the parameter-conditional posterior probabilities $P(\lambda_N=i/X^m,B)$ (This is the probability that would be used for decision making with $B = B_0$ if B_0 were known) over the posterior density $p(B/X^m)$. This posterior density summarizes knowledge about B_0 in the first m observations. Both terms are generated iteratively in the next section.

2.3 ITERATIVE GENERATION OF THE POSTERIOR DENSITIES

In this section, the key posterior densities in the decision rule derived in Sec. 2.2 are generated iteratively. First, $P(\lambda_N=i/X^m,B)$ is generated for $m = N$, $m > N$ and $m < N$. Then $p(B/X^m)$ is treated.

Case 1. $m = N$

From the Bayes rule,

$$P(\lambda_m=i/X^m,B) = \frac{p(X_m/\lambda_m=i,B,X^{m-1})P(\lambda_m=i/X^{m-1},B)}{p(X_m/X^{m-1},B)} \tag{2.3.1}$$

where the three terms on the right hand side can be accounted for in the following three steps.

1. $p(X_m/\lambda_m=i,B,X^{m-1})$ is given by (2.1.2)
2. By the total probability law

$$P(\lambda_m=i/X^{m-1},B) = \sum_{j=1}^{M} P(\lambda_m=i/\lambda_{m-1}=j,X^{m-1},B)P(\lambda_{m-1}=j/X^{m-1},B) \tag{2.3.2}$$

where (2.1.2) and Assumption 1 of Sec. 2.1 imply

$$P(\lambda_m=i/\lambda_{m-1}=j,X^{m-1},B) = p_{ji} \tag{2.3.3}$$

and $P(\lambda_{m-1}=j/X^{m-1},B)$ is available from the previous step of the iteration scheme.

3. Again by the total probability law and (2.1.2)

$$p(X_m/X^{m-1},B) = \sum_{i=1}^{M} f(X_m/B_i)P(\lambda_m=i/X^{m-1},B) \tag{2.3.4}$$

Case 2. $m > N$

Using arguments similar to those used in the previous development, it follows that

$$P(\lambda_N=i/X^m,B) = \frac{p(X_{N+1}\cdots X_m/\lambda_N=i,X^N,B)P(\lambda_N=i/X^N,B)}{p(X_{N+1}\cdots X_m/X^N,B)} \tag{2.3.5}$$

where $p(\lambda_N=i/X^N,B)$ is known from the previous iteration using steps 1, 2, 3 above and with m replaced by N

$$P(X_{N+1}\cdots X_m/X^N,B) = \sum_{i=1}^{M} p(X_{N+1}\cdots X_m/\lambda_N=i,X^N,B)P(\lambda_N=i/X^N,B) \tag{2.3.6}$$

The iteration procedure would be complete if $p(X_{N+1}\cdots X_m/\lambda_N=i,X^N,B)$ could be determined. This factor should be recognized as the heart of the look-ahead procedure that results from using future observations. To generate this last factor it is convenient to define $Z^N = [X_{N+1}\cdots X_m]$. Then, by (2.1.2),

$$p(Z^N/\lambda_N=i,X^N,B) = p(Z^N/\lambda_N=i,B) \tag{2.3.7}$$

But

$$\begin{aligned} p(Z^N/\lambda_N=i,B) &= \sum_{j=1}^{M} p(Z^N/\lambda_N=i,\lambda_{N+1}=j,B)P(\lambda_{N+1}=j/\lambda_N=i,B) \\ &= \sum_{j=1}^{M} p(Z^N/\lambda_N=i,\lambda_{N+1}=j,B)p_{ij} \end{aligned} \tag{2.3.8}$$

Similarly,

$$p(Z^N/\lambda_N=i,\lambda_{N+1}=j,B) = \sum_k p_{jk} p(Z^N/\lambda_N=i,\lambda_{N+1}=j,\lambda_{N+2}=k,B) \qquad (2.3.9)$$

Hence this procedure can be repeated until

$$p(Z^N/\lambda_N=i\ldots\lambda_{m-1}=t,B) = \sum_q p_{tq} p(Z^N/\lambda_N=i,\lambda_{N+1}=j,\ldots,\lambda_m=q,B) \qquad (2.3.10)$$

where

$$p(Z^N/\lambda_N=i,\lambda_{N+1}=j,\ldots,\lambda_m=q,B) = f(X_{N+1}/B_j)\ldots f(X_m/B_q) \qquad (2.3.11)$$

Although the above procedure shows the decision rule to be a fixed-memory rule (the number of observations stored is fixed) the memory increased linearly with $m-N$ and the number of computations in (2.3.8) to (2.3.11) increases exponentially with $m-N$. Hence, the look-ahead is expensive.

Case 3. $m < N$

Past samples are being used to make decision about future states. For $m < N$

$$P(\lambda_N=i/X^m,B) = \sum_{j=1}^{M} P(\lambda_N=i/\lambda_m=j,X^m,B) P(\lambda_m=j/X^m,B) \qquad (2.3.12)$$

But, by (2.1.2),

$$P(\lambda_N=i/\lambda_m=j,X^m,B) = P(\lambda_N=i/\lambda_m=j) \qquad (2.3.13)$$

which is the $(j,i)^{th}$ element of P^{N-m} and $P(\lambda_m=j/X^m,B)$ can be obtained iteratively using steps 1, 2, 3 above. In this case, the number of computations involved in making decisions on future states increases linearly with $m-N$.

Finally, an iterative form for $p(B/X^m)$ is established.

$$p(B/X^m) = \frac{p(X_m/X^{m-1},B)\,p(B/X^{m-1})}{p(X_m/X^{m-1})} \tag{2.3.14}$$

where $p(X_m/X^{m-1},B)$ is given by (2.3.4), $p(B/X^{m-1})$ is available from the previous step and

$$p(X_m/X^{m-1}) = \int p(X^m/X^{m-1},B)\,p(B/X^{m-1})\,dB \tag{2.3.15}$$

2.4 ANALYSIS OF THE ITERATIVE PROCEDURE

In this section, a special but important case of the previous decision rule is studied. The iterative structure of the rule is investigated and interpreted. The start of the iterative procedure is illustrated and the problems encountered in machine implementation of the rule are discussed.

If $m=N$ and a 0-1 loss function is used then the result is the minimum probability of error rule, δ_e, given by (2.2.11), for determining the present state of the POMS, described in Sec. 2.1, on the basis of past and present observations. This rule is summarized below.

$$\delta_e(X^N) = i \quad \text{if} \quad P(\lambda_N=i/X^N) \geq P(\lambda_N=j/X^N) \quad \forall\, i \neq j \tag{2.4.1}$$

where, from (2.2.6),

$$P(\lambda_N=i/X^N) = \int P(\lambda_N=i/X^N,B)\,p(B/X^N)\,dB \tag{2.4.2}$$

with, from (2.3.1) and (2.3.2),

$$P(\lambda_N=i/X^N,B) = \frac{f(X_N/B_i)\,P(\lambda_N=i/X^{N-1},B)}{p(X_N/X^{N-1},B)} \tag{2.4.3}$$

and

$$P(\lambda_N=i/X^{N-1},B) = \sum_{j=1}^{M} p_{ji} P(\lambda_{N-1}=j/X^{N-1},B) \tag{2.4.4}$$

and, from (2.3.14),

$$p(B/X^N) = \frac{p(X_N/X^{N-1},B)\,p(B/X^{N-1})}{p(X_N/X^{N-1})} \tag{2.4.5}$$

All the terms in (2.4.2)-(2.4.5) are either known, available from the previous step or can be calculated from those given above.

The rule δ_e is a fixed memory, iterative, optimal decision rule. The parameter-conditional state posterior probability, $P(\lambda_N=i/X^N,B)$, is the state posterior probability given by (1.2.3) (where B_0 was assumed known) as a function of the unknown parameter. Equations (2.4.3) and (2.4.4) generate this term iteratively in a manner similar to that of Sec. 1.2 but conditioned on knowledge of the unknown parameter. That is, this probability at time N-1 is projected with the transition matrix in (2.4.4) to $P(\lambda_N=i/X^{N-1},B)$ which is used in the Bayes rule in (2.4.3) in incorporate the information provided by X_N and to generate $P(\lambda_N=i/X^N,B)$. Everything in the procedure is conditioned on knowledge of B_0 and information about B_0 is summarized by $P(B/X^N)$ which is generated iteratively in (2.4.5) and is introduced into the decision rule by the averaging procedure given in (2.4.2).

The starting procedure for the iterative scheme is given below. At step 1

$$P(\lambda_1=i/X^1,B) = \frac{f(X_1/B_i)P(\lambda_1=i)}{p(X_1/B)} \tag{2.4.6}$$

where

$$p(X_1/B) = \sum_{i=1}^{M} f(X_1/B_i)P(\lambda_1=i) \qquad (2.4.7)$$

then

$$p(B/X_1) = \frac{p(X_1/B)p_0(B)}{p(X_1)}$$

where

$$p(X_1) = \int p(X_1/B)p_0(B)dB$$

At step 2

$$P(\lambda_2=i/X_1,X_2,B) = \frac{f_i(X_2/B_i)P(\lambda_2=i/X_1,B)}{p(X_2/X_1,B)}$$

where

$$p(X_2/X_1,B) = \sum_{i=1}^{m} f_i(X_2/B_i)P(\lambda_2=i/X_1,B)$$

and

$$P(\lambda_2=i/X_1,B) = \sum_{j} p_{ji}P(\lambda_1=j/X_1,B)$$

then

$$p(B/X_1,X_2) = \frac{p(X_2/X_1,B)p(B/X_1)}{p(X_2/X_1)}$$

where

$$p(X_2/X_1) = \int p(X_2/X_1,B)p(B/X^1)dB$$

this procedure can be repeated up to time N.

Despite the desirable features of the above rule it has one major drawback. At each step, m+1 functions of B must be stored. If this rule is to be machine implemented it can only be done under one of the following conditions which are characteristic of general Bayesian learning.

1. The parameter vector B is a discrete random variable taking on a finite number of values. That is, B_0 is known to be one of a finite number of values. This allows storage of the required densities but is a highly restrictive assumption.
2. The parameter vector B is a continuous random variable but a finite dimensional sufficient statistic is available for the unknown parameter. Under these conditions, only a function of the observations need to be stored [S-5]. However, such statistics do not usually arise in unsupervised learning problems because the class of densities involved are mixtures.
3. In general, the only way to make use of the rule is by quantization of the parameter space, thus reducing the problem to case 1 above. By quantizing fine enough it is possible to get arbitrarily close to the optimum solution at the expense of increased memory [F-1]. However, the memory grows exponentially with the dimension of the parameter vector B, making the method feasible only for problems with a small number of parameters. An example is given in Sec. 3.6 which indicates the extent of the storage limitations.

2.5 THE LEARNING FEATURE OF THE OPTIMUM DECISION RULE

In this section, some limiting properties of the posterior densities that comprise the optimum decision rule (2.2.10) are discussed in a manner that exhibits the learning capability of the rule. First it is shown that the state posterior probability $P(\lambda_N=i/X^m)$ given by (2.2.6) with N fixed converges with probability one as the number of observations is increased. Then, the conditions under which

$$p(B/X^m) \xrightarrow{m} \delta(B-B_0) \qquad \text{wp 1} \tag{2.5.1}$$

where $\delta(B-B_0)$ is the Dirac delta function are stated. Equation (2.5.1) can be restated as saying that the posterior distribution of B, which summarizes all the information about B_0 contained in X^m, approaches wp1 a distribution whose mass is grouped about B_0 so that B_0 is learned. Then, for all practical purposes,

$$p(\lambda_N=i/X^m) \xrightarrow{m} p(\lambda_N=i/X^\infty, B_0) \tag{2.5.2}$$

and the rule adapts, or converges, to the optimal rule that would be obtained if B_0 were known.[11]

The statistical stability of $P(\lambda_N=i/X^m) = Y_m$ can be demonstrated by showing $\{Y_m\}$ to be a bounded martingale. Then, convergence follows immediately from a theorem of Doob which says that every bounded martingale converges with probability one [D-2]. To show Y_m to be a martingale it is sufficient to prove that

$$E[Y_{m+1}/X^m] = Y_m$$

But

$$\begin{aligned} E[Y_{m+1}/X^m] &= \int P(\lambda_N=i/X^{m+1})\, p(X^{m+1}/X^m)\, dX_{m+1} \\ &= \int \frac{p(X^{m+1}/\lambda_N=i)P(\lambda_N=i)}{p(X^{m+1})} \frac{p(X^{m+1})}{p(X^m)}\, dX^{m+1} \\ &= \frac{p(X^m/\lambda_N=i)P(\lambda_N=i)}{p(X^m)} = Y_m \end{aligned}$$

Since for each m Y_m is a probability,

$$|Y_m| \leq 1 \qquad \forall\ m$$

---

[11] The implied interchange of the limit and integration process has been carried out formally and (2.5.2) has not been proven for any mode of convergence.

and Y_m is bounded. Therefore $\lim_{m\to\infty} Y_m$ exists a.e. in the space of sequences $\{X^\infty\}$.

The conditions under which the parameter posterior density approaches a delta function are well known. Spragins [S-5] and Braverman [B-3] have given sufficient conditions for (2.5.1) to hold through an interpretation of the 0-1 law of probability [L-1]. These conditions are presented below.

1. $p(B/X^N)$ is computed using the Bayes rule.
2. $p_0(B)$ is positive in some sphere about B_0
3. There exist functions $\{f_m(X^m)\}_1^\infty$ of the observations such that

$$f_m(X^m) \xrightarrow{m} B_0 \qquad \text{wp1}$$

That these conditions are satisfied for decision rule (2.2.10) is now demonstrated. Condition 1 follows from (2.3.7). Condition 2 is assumed. Condition 3 is the major requirement and can be restated as saying that a strongly consistent estimator for B_0 must be exhibited. In Chapter III, a variety of such estimators are established by placing constraints on the transition matrix and the family of component densities.

2.6 RELATED INFERENCE PROBLEMS

Some of the properties of the previously-derived decision rule can be used to solve additional decision problems of interest. For example, in the adaption process, B_0 is learned but estimates of B_0 are never actually generated. Since the parameter posterior density converges wp1 to a dirac delta at B_0, any property of this density,

such as the mean, maximum or median, converges wpl to B_0 also. Furthermore, it is well known that the mean of the posterior density is the minimum mean-square estimate of B_0 and hence can be interpreted as an optimal estimate for B_0.

Another decision problem of interest is that of determining which class or subset of the states the system is in. In particular, if the transition matrix P is block diagonal with q blocks, the states of the system are divided into q classes with the property that the system stays in the class in which it starts.[12] Then, if $P(w_i/X^m)$ is the conditional probability that the system is in class i,

$$P(w_i/X^N) = \sum_{j\in\gamma_i} P(\lambda_N=j/X^N) \qquad i = 1,2,\ldots,q \tag{2.6.1}$$

where γ_i is that subset of the state space corresponding to class i. The probability $P(\lambda_N=j/X^m)$ can be generated iteratively in the manner of Sec. 2.3 and an optimal decision rule for choosing the system class on the basis of X^N is to pick the class for which the class posterior probability $P(w_i/X^m)$ is largest. In order to exhibit some of the properties of the above decision rule, (2.6.1) is rewritten as

$$P(w_i/X^N) = \frac{p(X_N/w_i,X^{N-1})P(w_i/X^{N-1})}{p(X_N/X^{N-1})} \tag{2.6.2}$$

where

$$p(X_N/X^{N-1}) = \sum_{i=1}^{q} p(X_N/w_i,X^{N-1})P(w_i/X^{N-1}) \tag{2.6.3}$$

and

---

[12] The corresponding Markov chain is said to have q noncommunicating classes.

$$p(X_N/w_i,X^{N-1}) = \sum_{j\in\gamma_i} \int p(X_N/\lambda_N=j,w_i,B^i,X^{N-1})P(\lambda_N=j/w_i,X^{N-1},B^i)$$
$$p(B^i/w_i,X^{N-1})dB^i \quad (2.6.4)$$

with B^i the parameter vector for the component densities in the i^{th} class and $\{P_0(w_i)\}_1^q$, the prior class probabilities, assumed known. Then (2.6.2) indicates that the class posterior probabilities can be generated using the Bayes rules with normalizing factor given by (2.6.3). All the terms in the mixture (2.6.4) are conditioned on knowledge of the class and thus can be generated iteratively using only the block of the transition matrix and component densities corresponding to the given class. By an appropriate interpretation of Spragins' conditions for convergence, it follows that if $P_0(w_{i_0}) > 0$ and there exists a strongly consistent estimator for i_0 then

$$p(w_j/X^m) \xrightarrow{m} \Delta_{ji_0} \qquad \text{wp1} \qquad (2.6.5)$$

where i_0 is the true class. Conditions for such an estimator to exist are discussed in Sec. 3.8.

Finally, the above two decision problems can be combined into that of estimating the unknown parameters defining the class in which the system is. If the parameter vectors are the same dimension for each class, the optimal estimate is given by the mean of

$$p(B/X^N) = \sum_{i=1}^{q} p(B/w_i,X^N)P(w_i/X^N) \qquad (2.6.6)$$

This procedure is referred to as adaptive estimation and conditions for convergence are given in Sec. 3.8.

2.7 CONCLUSIONS

Optimal decision making for a POMS with unknown parameters in the component densities has been the topic of this chapter. Assuming a prior distribution over the parameter space, an optimal decision rule was defined in Sec. 2.1 to be that rule in a given class of rules which minimizes the Bayes risk (2.1.3). This class of rules, similar to that of Sec. 1.2 where the component densities were assumed known, includes rules with the capability of using some future observations (look-ahead mode) as well as past observations to make decision about the state of the system at a given time.

The optimal decision rule (2.2.10) was derived in Sec. 2.2 and its basic components (2.2.6)-(2.2.8) were generated iteratively in Sec. 2.3. It was shown that for the look-ahead mode the memory grows linearly and the number of computations exponentially with the number of future samples used. However, the extent to which future observations affect the risk needs investigation. It is clear from the derivation that these results can be extended to include time-varying transition probabilities.

As emphasized in Sec. 1.4, the optimal decision rule is a fixed-memory, iterative rule with a structure similar to that of the rules in Sec. 1.2. In general, only a high storage quatization procedure can be used to implement the rule. This suggests that the main use of the optimal rule may be to evaluate related suboptimal, low storage rules in the hope that conclusions can be extrapolated to the more complicated cases.

In Sec. 2.5, it was shown that, under the stated conditions, the parameter posterior density (2.2.7), which summarizes the knowledge about B_0, converges to a dirac delta function. Thus, B_0 is learned and the rule adapts.

Finally, in Sec. 2.6, various results established in the previous sections were used to solve related inference problems. In particular, a class of estimators for B_0 was established including an optimal (minimum mean square) estimator. An optimal decision rule was constructed for determining to which class the state of the system belongs among a set of noncommunicating classes of states. An optimal adaptive estimator was constructed for the parameters in the component densities associated with the active class of states. These examples indicate the versatility of the decision problem of Sec. 2.1.

CHAPTER III

CONSISTENT ESTIMATORS

When the component densities in a Partially Observable Markov System (POMS) are specified to within a parameter vector, but no prior distribution for the parameter is available, optimal decision making as defined in Sec. 2.1 is no longer possible. However, the observations still contain information about B_0, the true but unknown value of the parameter. If a strongly-consistent estimator for B_0 (a function of the observations that converges with probability one to B_0) can be found, decision rules can be constructed by treating the estimate at a given time as if it were the true value. The decision rule of Sec. 1.2, where the component densities were assumed known, can then be used. Hopefully, as more observations are taken, decision rules constructed in this manner adapt or converge to the optimal rule, which uses the true component densities.[13] Such rules, unlike those of Chapter II, may not extract information about B_0 in an optimal manner but do approach the optimum as more observations are taken.

There are several other reasons for investigating consistent estimators, some of which are listed below.

1. Even when a prior distribution on the parameters is available suboptimal rules may be easier to implement than the high-storage quantization procedure of Sec. 2.4 implicit in the optimal rule.[14]

2. In Sec. 2.5, it was shown that a strongly-consistent estimator for B_0 must be exhibited to ensure that this parameter

---

[13] This convergence problem is discussed in Chapter V.

[14] Optimal and suboptimal rules are compared with regard to implementation in Chapter V.

will be learned during the operation of the optimal rule. Hence conditions for the existence of such estimators are important.

3. In some applications the true parameter values may be of interest in themselves. The estimators suggested in Sec. 2.6 not only require parameter prior distributions but suffer from the implementation difficulties mentioned in 1 above.

This chapter deals mainly with the problem of finding strongly-consistent estimators for the parameters defining the component densities of a POMS. In Sec. 3.1 the estimation problem is defined and some important assumptions are discussed. The properties of the observation process are listed in Sec. 3.2 while estimation tools are developed in Sec. 3.3. In Sec. 3.4 and 3.5, estimation techniques developed for the case of independent identically distributed (i.i.d.) observations discussed in Sec. 1.3 are extended to the more general, dependent, nonstationary case under study in this chapter. Section 3.6 provides computer-simulated results for an example illustrating some of these estimation techniques. Generalizations of the problem defined in Sec. 3.1 are discussed in Sec. 3.7 and 3.8. Finally, Sec. 3.9 summarizes the main results of the chapter.

3.1 THE ESTIMATION PROBLEM

The problem of finding a strongly-consistent estimator for B_0, the unknown parameter in a set of component distributions, is precisely stated and some of the assumptions necessary to ensure the existence of such estimators are discussed in this section.

A POMS similar to that of Sec. 2.1 is considered.[15] The state activity is described by a transition matrix P and initial probability state vector p_1. The distribution of the observed process $\{X_N\}_1^\infty$ is defined by a set of distinct univariate component CDF's $\{F_i(\cdot)\}_1^M$ corresponding to the component densities of Sec. 1.1; $F_i(\cdot)$ is assumed to be an unknown element of the family $\mathfrak{F} = \{F(X;\alpha)\}_{\alpha\in A}$ indexed by a point α in a subset A of the real line.[16] As indicated in Sec. 1.1, X_N, the random variable observed at time N, has CDF $K_N(X)$ which is an element of the set of mixtures $H_N = \{K_N(X;B)\}_{B\in\mathcal{B}'}$, where

$$K_N(X;B) = \sum_{i=1}^{M} F(X;B_i)P(\lambda_N=i) \tag{3.1.1}$$

In (3.1.1) $B = [B_1,B_2,\ldots,B_M]$ and $\mathcal{B}'$ is the set of all such vectors with distinct components $B_i \in A$. That is $\mathcal{B}'$ is a subset of M-dimensional Euclidean space. Then there exists at least one point $B_0 \in \mathcal{B}'$ such that

$$K_N(X) = K_N(X;B_0) \qquad \forall\, N \tag{3.1.2}$$

The problem is to find $B_m(X^m)$, a function of the observations $X^m = [X_1,\ldots,X_m]$, such that

$$P(B_m(X^m) \overset{m}{\to} B_0) = 1 \tag{3.1.3}$$

The major assumptions necessary for the construction of such estimators are listed below. They are assumed to hold throughout the chapter and will be referred to when needed.

A-1. The observed process $\{X_i\}_1^\infty$ is state-conditionally independent.[17]

---

[15] Here B is no longer being treated as a random variable.

[16] Scalar observations and parameters are assumed for simplicity but all ideas extend to the vector case.

[17] This corresponds to the assumption of state-parameter conditional

A-2. For each X, $F(X;\cdot)$ is a continuous function on A and for each α, $F(\cdot;\alpha)$ is continuous in X; $i = 1,2,\ldots,m$.

A-3. B_0 is an interior point of $\mathcal{B}$ and $\mathcal{B}$ is a compact subset of $\mathcal{B}'$.

A-4. $\{\lambda_N\}_1^\infty$ is a regular Markov chain.[18]

Assumption A-1 was introduced in Sec. 1.1 and is repeated here for convenience. It is a key assumption in developing the estimation tools of Sec. 3.3. Assumptions A-2 and A-3 are standard requirements for developing the existence and convergence properties of estimators.

Assumption A-4 characterizes the class of POMS to which the techniques of this chapter apply. It ensures that the observations contain information about all the components of B_0. More explicitly, a regular Markov chain has the property that it is possible to be in any state after some finite number of steps no matter what the starting state [K-4]. Hence, among a large number of observations there will be a large number of representatives from each of the component distributions. As shown in Sec. 3.2, this assumption also implies the asymptotic ergodicity of the observed process, which is the basis of the estimation strategy to come.

In addition to assumptions A-1 through A-4 listed above, a uniqueness condition on B_0 is required. This is a standard assumption in estimation problems and takes different forms depending on the method used. For the estimation problem defined above a special uniqueness condition arises since, in general, mixtures do not have a unique decomposition into allowable component distributions. So (3.1.2) may

---

[18] A regualr Markov chain is one that has no transient states and only one ergodic class with no cyclically moving sub-classes.

not define B_0 uniquely. To avoid this type of ambiguity the mapping defined by (3.1.1) from $\mathcal{B}$ onto a subset of H_N, (for a fixed family of component distributions) must be one-to-one. This subset of H_N is then an example of an identifiable class of parameter-indexed mixtures [H-2]. The concept of identifiability is discussed in the Appendix where it is shown that this property, which is necessary for all the methods of this chapter, can be ensured by placing constraints on the family of component distributions $\mathfrak{F}$ and the parameter space $\mathcal{B}'$.

3.2 PROPERTIES OF THE OBSERVATIONS

In order to establish estimation strategies it is necessary to study some properties of the observation process $\{X_i\}_1^\infty$. In this section, properties of the system state activity and the resulting observations are developed by making use of assumptions (A-1) and (A-4).

When $\{\lambda_i\}_1^\infty$ is a regular Markov chain with transition matrix P and initial probability state vector $\underline{p}_1$, it has the following three properties [K-4].

1. There exists a unique vector $\underline{\pi} = [\pi_1, \ldots, \pi_M]$ such that

$$\underline{\pi} = \underline{\pi} P \tag{3.2.1}$$

2. If $\underline{p}_1 = \underline{\pi}$, $\{\lambda_i\}_1^\infty$ is a stationary process.

3. $$\underline{p}_N = \underline{p}_{N-1} P = \underline{p}_1 P^{N-1} \xrightarrow{N} \underline{\pi} \qquad \forall\, \underline{p}_1 \tag{3.2.2}$$

The vector $\underline{\pi}$ is called the stationary probability state vector. Given P, (3.2.1) can usually be solved directly for $\underline{\pi}$. Otherwise (3.2.2) suggests a convenient algorithm for approximating $\underline{\pi}$.

Properties 2 and 3 indicate that a regular Markov chain is asymptotically stationary. Such activity in a POMS suggests that the

observations might exhibit some form of asymptotic stability. In fact, $\underline{p}_N$ is the set of mixing parameters for the mixture $K_N(X)$ which governs the observations at time N. Hence properties 1 and 3 imply there exists a unique mixture cdf $K_\pi(X) = \sum_1^M F_i(X)\pi_i$ such that

$$\sup_{-\infty<X<\infty} \left| K_N(X) - K_\pi(X) \right| \leq \sum_{i=1}^{M} \left| \pi_i - P(\lambda_N=i) \right| \xrightarrow{N} 0 \qquad (3.2.3)$$

Hence the cdf of X_N, the random variable observed at time N, uniformly approaches a unique limit mixture; $K_\pi(X)$ is the 1^{st} order distribution of the stationary process which results when $\underline{p}_1 = \underline{\pi}$.

To gain further insight into the statistical structure of the observation process, two results proved by Raviv [R-2] are presented below in slightly modified form.

Lemma 3.2.1 Under assumption (A-1) each random variable in the process $\{X_i\}_1^\infty$ can be represented as

$$X_N = \phi(M_N)$$

where $\phi(\cdot)$ is a Baire function and $\{M_i\}_1^\infty$ is a Markov process satisfying Doeblins Hypothesis.[19]

Lemma 3.2.2 Under assumptions (A-1) and (A-2), if $\underline{p}_1 = \underline{\pi}$ then $\{M_i\}_1^\infty$ and $\{X_i\}_1^\infty$ are stationary and ergodic (metrically transitive) processes.

The above lemmas and property 3 imply that the observation process is asymptotically ergodic for a class of POMS defined by assumptions A-1 and A-4. Furthermore, for this class of systems the limiting behavior of the observations is independent of the initial system activity.

---

[19] Doob [D-2], pg. 192 .

3.3 THE BASIC TOOLS OF ESTIMATION

When the observation process in an estimation problem is a sequence of independent, identically-distributed (i.i.d.) random variables, the key tools for constructing estimators are the Law of Large Numbers (LLN) and the Glivenko-Cantelli Theorem (GCT). If the observation process of Sec. 3.1 is ergodic, the ergodic theorem establishes appropriate extensions of these tools. The fact that the observation process for the class of estimation problems considered in this chapter is asymptotically ergodic suggests that these important tools might be extended to this case. In this section the needed generalization of the LLN and the GCT are established and a basic estimation strategy is stated for the problem of Sec. 3.1.

Using lemma 3.2.1 and assumption A-4, Raviv [R-2] has essentially proved the following theorem for the observation process $\{X_i\}_1^\infty$ of a POMS with arbitrary initial probability state vector $\underline{p}_1$.[20]

THEOREM 3.3.1 If $g(\cdot)$ is a Baire function integrable with respect to $K_\pi(X)$ then

$$\frac{1}{N}\sum_1^M g(X_i) \xrightarrow{N} E_{\pi_0} g(X) \qquad \text{wp1} \tag{3.3.1}$$

where

$$E_{\pi_0} g(X) = \sum_1^M \pi_i \int g(X)\, dF_i(X) \tag{3.3.2}$$[21]

Theorem 3.3.1 is the required extension of the LLN. It provides a means for establishing strongly-consistent estimators for any expectation of $K_\pi(X)$.

---

[20] The proof which consists of a direct application of Theorem 6.2 of Doob appears as part of the proof of Raviv's Lemma 2.5.

[21] E_{π_0} will be used to denote expectation with respect to the true limit mixture of the POMS; i.e. the distribution with cdf $K_\pi(X) = K_\pi(X;B_0)$.

The problem of estimating $K_\pi(X)$ from the observations $\{X_i\}_1^\infty$ is now considered and an extension of the GCT is presented.

The function

$$\hat{K}_N(x) = 1/N \sum_{k=1}^{N} I_{[r: r \le x]}(X_k) \tag{3.3.3}$$

where $I_A(\cdot)$ is the indicator function on the set A is called the empirical distribution function; $\hat{K}_N(x)$ is the proportion of samples from the set $X^N = [X_1, \ldots, X_N]$ that are less than or equal to x. Since

$$E_{\pi_0} I_{[r: r \le x_0]}(X) = K_\pi(x_0) \le 1$$

and[22]

$$E_{\pi_0} I_{[r: r \le x_0]}(X) = K_\pi(x_0 - 0) \le 1$$

the following lemma is an immediate consequence of Theorem 3.3.1.

Lemma 3.3.1 For every real x

$$P\{\hat{K}_N(x) \xrightarrow{N} K_\pi(x)\} = 1$$

$$P\{K_N(x-0) \xrightarrow{N} K_\pi(x-0)\} = 1$$

Lemma 3.3.1 is the key step in extending the GCT.

THEOREM 3.3.2

$$P\{\sup_{-\infty < x < \infty} |K_N(x) - K_\pi(x)| \xrightarrow{N} 0\} = 1$$

With Lemma 3.3.1 the proof is exactly the same as that of the classical theorem [T-2] and will be omitted.

Theorems 3.3.1 and 3.3.2 lead to a general estimation strategy for determining B_0. As indicated by the discussion of Sec. 3.1 and

---

[22] $f(x-0) = \lim_{\epsilon \to 0} f(x-\epsilon)$; $\epsilon > 0$.

3.2 the unknown limit mixture $K_{\pi}(X;B_0)$ is an element of the class of mixtures $H_{\pi} = \{K_{\pi}(X;B)\}_{B\in\mathcal{B}}$ where

$$K_{\pi}(X;B) = \sum_{1}^{M} F(X;B_i)\pi_i$$

The main idea is to <u>treat the process $\{X_i\}_1^{\infty}$ as if it were a sequence of independent random variables with common CDF $K_{\pi}(X;B_0)$</u>. The schemes developed for the i.i.d. case are then used with the classical convergence tools replaced by Theorems 3.3.1 and 3.3.2. This procedure is illustrated in Sec. 3.4 and 3.5. The general format is to establish an estimator, present a convergence theorem and evaluate the results. Assumptions A-1 and A-4 will be assumed through the rest of the chapter.

3.4 CLASSICAL ESTIMATION METHODS

In this section, the principle behind the classical method of moments and the maximum likelihood method are used to solve the estimation problem defined in Sec. 3.1. Both methods involve the use of sample averages whose convergence is guaranteed by Theorem 3.3.1.

The method of moments has been used frequently for mixture resolving [B-2][C-5][P-3][R-1][R-4] and is treated in the most general form by Chien and Fu [C-2]. The procedure requires equating M sample moments to the corresponding population moments of $K_{\pi}(X;B)$. A solution, if one exists, of the resulting equation in B can then be taken as an estimate for B_0. More specifically, if F(B) is a vector of population moments of $K_{\pi}(X;B)$, which are assumed to exist, and T_m is the corresponding set of sample moment constructed from the observation X^m, then by Theorem 3.3.1

$$T_m \xrightarrow{m} T_0 \qquad \text{wp1} \tag{3.4.1}$$

where T_0 is a vector of moments of the true limit mixture $K_\pi(X;B_0)$. The set of existing solutions of

$$T_m = F(B) \tag{3.4.2}$$

may contain candidates for strongly-consistent estimates of B_0. In order to obtain a more explicit result the following uniqueness condition is assumed. The function $F(\cdot)$ is a one-to-one mapping from $\mathcal{B}$ onto τ, a subset of M-dimensional Euclidean space. Under this condition, a consistent estimator can be constructed as indicated by the following theorem.

THEOREM 3.4.1 If the above uniqueness condition is satisfied then, with probability one, there exists a sequence of solutions of (3.4.2), say $\{B_m\}_{m_0}^{\infty}$, which converges to B_0.

Proof. Assumption A-2 implies that $F(\cdot)$ is a continuous function on $\mathcal{B}$. Hence by the uniqueness condition $F^{-1}(\cdot)$ is defined and continuous [R-5]. Equation (3.4.1) and the fact that B_0 is an interior point of $\mathcal{B}$ imply that for every sequence of observations, except possibly those in a set of measure zero, there exists an m_0 such that $T_m \in \tau$ for $m > m_0$. Thus for $m > m_0$, $B_m = F^{-1}(T_m)$ and converges to B_0.

In general (3.4.2) can have more than one solution and prior information concerning the particular application must be used to select a convergent sequence of solutions. This corresponds to restricting $\mathcal{B}'$ to $\mathcal{B}$ as given in the uniqueness condition.

Equation (3.4.2) can pose some difficult computation problems since, in general, (3.4.2) is a set of complicated nonlinear equations and an iterative algorithm for finding a zero of a function must be used at each step. Since the sample moments can be generated iteratively, the storage requirement is fixed. When B can be found as an explicit function of T_m the resulting estimator is very desirable with respect to implementation.

The maximum likelihood method has been used by Patrick [H-2] for mixture-resolving. He treated a mixture of Gaussian distributions in detail. The version of maximum likelihood estimation given here is essentially a modification of that given by Wilks [W-1].

The following notation is introduced.

$$S_j(X;B) = \frac{\partial \text{ Log } dK_\pi(X;B)}{\partial B_j} \quad j = 1,2,\ldots,M \tag{3.4.3}$$

$$\underline{S}(X;B) = [S_1(X;B),\ldots,S_M(X;B)] \tag{3.4.4}$$

$$A_j(B;B') = \int S_j(X;B)\, dK_\pi(X;B') \tag{3.4.5}$$

$$\underline{A}(B;B') = [A_1(X;B),\ldots,A_M(X;B)] \tag{3.4.6}$$

Now, $K_\pi(X;B)$ is said to be regular in $\mathcal{B}$ with respect to its first derivative if

$$E_\pi S_j(X;B) \equiv A_j(B;B) = \frac{\partial}{\partial B_j} \int dK_\pi(X;B) \equiv 0 \;\forall\; B \in \mathcal{B},\ j = 1,2,\ldots,M \tag{3.4.7}$$

Under this condition Theorem 3.3.1 implies

$$\underline{S}(X^m;B) \equiv 1/m \sum_1^m \underline{S}(X_i;B) \xrightarrow{m} A(B;B_0) \tag{3.4.8}$$

Since $A(B_0,B_0) = 0$ a reasonable strategy for generating a consistent estimator is to choose among the roots of

$$\underline{S}(X^m;B) = 0 \tag{3.4.9}$$

This idea is made more explicit in the following theorem.

THEOREM 3.4.2 If $K_\pi(X;B)$ is regular in $\mathcal{B}$ with respect to its first derivative, then there exists a sequence of solutions of (3.4.9) which converges with probability one to B_0. In particular, if (3.4.9) has a unique solution B_m for $m > m_0$ then the sequence $\{B_m\}_{m_0}^{\infty}$ converges with probability one to B_0.

The proof, which is in the same spirit as that given for the method of moments and which follows from (3.4.8), is analogous to that given by Wilks and will be omitted. The same comments can be made about the uniqueness condition and the computational problem of finding roots of (3.4.9) as were mentioned for the method of moments. However in this case, all the samples must be stored and the form of (3.4.9) changes at each step. The same problem was faced in the i.i.d. case [K-2].

The interpretation of the above method as a maximum likelihood method is obvious if the likelihood function is taken to be the product $K_\pi(X_1;B),\ldots,K_\pi(X_m;B)$. Then, under suitable restrictions, the value of B which maximizes the likelihood function satisfies (3.4.9).

3.5 MINIMUM DISTANCE ESTIMATION METHODS

In this section, the minimum distance principle is used to construct estimators for B_0. Given an appropriate distance measure $\rho(\cdot,\cdot)$ and the empirical distribution function for $K_\pi(X;B_0)$, $\hat{K}_N(X)$

defined in (3.3.3), and estimator B_m can be defined as an element in $\mathcal{B}$ which minimizes $\rho(\hat{K}_m(X), K_\pi(X;B))$. Theorem 3.3.2 is the main tool for establishing convergence. This method has been used for mixture resolving by Patrick [H-2], Stewart [H-3], Choi [C-3], and Deely and Kruse [D-1] and their ideas apply to the problem.

Patrick, Stewart, and Deely used the sup norm for a distance measure.[23] Analogously, the estimate B_m is defined by

$$\min_{B\in\mathcal{B}} \|\hat{K}_m(X) - K_\pi(X;B)\| = \|K_m(X) - K_\pi(X;B_m)\| \tag{3.5.1}$$

The existence of B_m is ensured by A-2 and A-3; i.e. B_m is obtained by minimizing a continuous function over a compact set. The following theorem establishes the consistency of such estimates.

THEOREM 3.5.1 If $H_\pi = \{K_\pi(X;B)\}_{B\in\mathcal{B}}$ is an identifiable class of mixtures, then B_m, defined by (3.5.1), converges with probability one to B_0.

Proof. If $Q_m(B) = \|\hat{K}_m(X) - K_\pi(X;B)\|$ then (3.5.1) and Theorem 3.3.2 imply

$$Q_m(B_m) \le Q_m(B_0) \xrightarrow{m} 0 \quad \text{wp1} \tag{3.5.2}$$

Since

$$\|K_\pi(X;B_0) - K_\pi(X;B_0)\| \le Q_m(B_m) + Q_m(B_0) \tag{3.5.3}$$

it follows that

$$K_\pi(X;B_m) \xrightarrow{m} K_\pi(X;B_0) \quad \text{wp1} \tag{3.5.4}$$

---

[23] $\|g(X)\| = \sup_{-\infty<X<\infty} |g(X)|$.

Assumption A-2 and A-3 and the fact that H_π is identifiable imply that the mapping defined by $K_\pi(X;B)$ from $\mathcal{B}$ to H_π is a one-to-one, continuous mapping from a compact set to a Hausdorf space. Hence the inverse exists and is continuous [R-5]. Thus (3.5.4) implies $B_m \xrightarrow{m} B_0$ wpl.

Another estimator analogous to that given by Choi can be constructed using the distance function

$$S_m(B) = \int (K_\pi(X;B) - \hat{K}_m(X))^2 d\hat{K}_m(X)$$

Then B_m is defined as the element of $\mathcal{B}$ which minimize $S_m(B)$. Under the same conditions as those in Theorem 3.5.1, the existence and strong consistency of B_m can be demonstrated from Theorem 3.3.2 and arguments analogous to those given by Choi.

The conditions needed to establish the existence of the above estimators are weakest considered in this thesis under which a strongly consistent estimator for B_0 can be established. Hence they are key assumptions which guarantee the learning property of the optimal rule given in Chapter II. Identifiability is a necessary condition for the uniqueness conditions required by the method of moments and the maximum likelihood method but, in general, is not sufficient.

The minimum-distance methods require storage of all samples and the use of a computational algorithm at each step. Such procedures have been developed for the i.i.d. case and can be applied here [C-3] [D-1].

3.6 AN EXAMPLE

In this section, a simple example of a POMS is used to illustrate some of the estimation techniques developed previously. Estimators are established from the method of moments, the maximum likelihood method and the optimal Bayes method of Sec. 2.6. To illustrate some basic properties of these estimators, results of computer simulation are presented.

The example consists of a POMS defined by the following.

regular[24] transition Matrix

$$P = \begin{bmatrix} .25 & .75 \\ .40 & .60 \end{bmatrix}$$

initial probability state vector

$$\underline{p}_1 = [.7 \quad .3]$$

component densities, Gaussian in form

$$f_1(X;B) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\sigma^{-2}(X+B)^2}$$

$$f_2(X;B) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\sigma^{-2}(X-B)^2}$$

The true mean B_0 is assumed to be in the interval (0,4). The stationary probability vector is easily calculated as

$$\underline{\pi} = [15/23 \quad 8/23]$$

and hence the limit mixture has the form

$$f_\pi(X;B) = \frac{1}{\sqrt{2\pi}\,\sigma}\left[15/23\ e^{-\frac{1}{2}\sigma^{-2}(X+B)^2} + 8/23\ e^{-\frac{1}{2}\sigma^{-2}(X-B)^2}\right] \qquad (3.6.1)$$

It is well-known that all moments of a Gaussian distribution exist and Patrick [H-2] has shown that a mixture of Gaussian distributions with

---

[24] As indicated in Chapter IV, if all the elements of a transition matrix are positive the corresponding Markov chain is regular.

unknown means is regular with respect to its first derivatives. Hence, the method of moments and the maximum likelihood method can be applied here.

A moment estimator is easily obtained using the second moment equation $E_{\pi}X^2 = B^2 + \sigma^2$. Then, Theorem 3.4.1 implies that the following estimator is strongly consistent.

$$\begin{aligned} B_m' &= \sqrt{S_m} && S_m \geq 0 \\ &= 0 && S_m < 0 \\ &= 2 && S_m > 4 \end{aligned} \tag{3.6.2}$$

where $S_m = 1/m \sum_{i=1}^{m} X_i^2 - \sigma^2$.

As indicated in Sec. 3.4, the maximum likelihood estimate B_m'' is defined as a solution to the following equation in the interval $[0,4]$

$$\sum_{i=1}^{m} \frac{-15(X_i+B)e^{-\frac{1}{2}\sigma^{-2}(X_i+B)^2} + 8(X_i-B)e^{-\frac{1}{2}\sigma^{-2}(X_i-B)^2}}{\sigma^2(15\, e^{-\frac{1}{2}\sigma^{-2}(X_i+B)^2} + 8\, e^{-\frac{1}{2}\sigma^{-2}(X_i-B)^2})} = 0 \tag{3.6.3}$$

Theorem 3.4.2 guarantees a sequence of roots that converges. Moreover, the computer solution[25] of (3.6.3) indicated only one root in $[0,4]$ for most values of m.

The Bayes estimator is obtained by quantizing the interval $[1,3]$ into 50 levels, .04 apart.[26] The estimate is defined by

$$B_m''' = \sum_{k=1}^{50} B_k p(B_k/X^m) \tag{3.6.4}$$

where B_k is the k^{th} quantization level and $p(B_k/X^m)$ is generated iteratively using equations (2.4.3)-(2.4.5) with B replaced by B_k.

---

[25] Under the conditions stated below.

[26] Additional prior information is assumed for Bayes estimator to emphasize its effect.

A uniform prior density is assumed. As pointed out in Sec. 2.6, if B_0 is in $[1,3]$ convergence is guaranteed by the existence of the previous two estimators.

The above estimators were implemented on a digital computer with a random number generator supplying the samples; B_0 was chosen equal to 2. For a given value of σ, $E(B_m)$ and $Var(B_m)$ were approximated for different values of m by averaging the results of twenty runs. On each run the same samples were used to generate estimates by all three methods. The results are presented in Table (3.6.1) for $\sigma = 1, 1.5, 2$ and $m = 25, 50, 100$.

For this example, Table (3.6.1) indicates some general trends. Not only do all the estimates converge on the average, but as more observations are taken the variances of the estimates decrease. Also, as the variance of the observations (controlled by σ) increases, the estimates become less accurate; i.e., $Var(B_m)$ and $|E(B_m) - B_0|$ increases. All three estimators seem to perform about the same with perhaps the Bayes estimator slightly better for small m and large σ. On the whole, the estimators behave much as they do for the i.i.d. case.

The above estimators can also be rated with regard to implementation. Since $\sigma^2 + S_m = (1+1/m)(\sigma^2 + S_{m-1}) + (1/m)X_m^2$, the moment estimator is a simple iterative one. The Bayes estimate is also iterative but, for this example, requires fifty times more storage and computation than the moment estimator. The maximum likelihood estimate is, by far, the worst requiring storage of all samples and use of a computational algorithm to find the zeros of (3.6.2) whenever an estimate is desired.

TABLE 3.6.1 COMPUTER SIMULATION RESULTS FROM 20 RUNS WITH $B_0 = 2$

		m = 25			m = 50			m = 100		
		MM	ML	B	MM	ML	B	MM	ML	B
$\sigma = 1$	$E(B_m)$	1.9494	2.9427	1.9414	1.9925	1.9997	1.9982	1.9984	2.0002	1.9983
	$Var(B_m)$	.0464	.0542	.0532	.0298	.0305	.0304	.0121	.0123	.0123
$\sigma = 1.5$	$E(B_m)$	2.0520	1.9961	1.9938	2.0536	2.0238	2.0099	1.9725	1.9369	1.9524
	$Var(B_m)$	.2291	.1693	.1266	.1089	.0998	.0994	.529	.0499	.0519
$\sigma = 2$	$E(B_m)$	1.8527	1.8279	1.9177	1.8767	1.9021	1.9139	1.9283	1.9474	1.9123
	$Var(B_m)$	.3493	.4566	.1498	.1495	.1223	.0959	.0833	.0608	.0627

Key

MM: Method of Moments

ML: Maximum Likelihood

B: Bayes

3.7 ALTERNATE STRATEGIES FOR THE ESTIMATION PROBLEM

In this section, three modifications of the basic estimation strategy outlined in Sec. 3.3 and illustrated in Sec. 3.4 and 3.5 are discussed briefly. Also, methods for handling extensions of the original estimation problem are indicated.

The first modification is one that can be used when π is unknown, as when P is unknown or $\underline{\pi}$ is too difficult to calculate. In such a case, $\underline{\pi}$ can be included in the set of unknowns of $K_{\pi}(X;B)$ and the methods of Sec. 3.4 and 3.5 can be used with an enlarged parameter vector.

The second modification is one that requires knowledge of P but computes $\underline{\pi}$ during the procedure. The main idea is to treat the random variables X^m as if they had common distribution $K_m(X)$, an element of $H_m = \{K_m(X;B)\}_{B\in\mathcal{B}}$. Then B_m, the m^{th} estimate, can be calculated using one of the principles of Sec. 3.4 and 3.5 with $K_{\pi}(X;B)$ replaced by $K_m(X;B)$. Since for each $B \in \mathcal{B}$, $K_m(X;B) \to K_{\pi}(X;B)$ uniformly in X, the estimate B_m should be consistent. As an example, a minimum distance estimator is considered; B_m is defined as the element in $\mathcal{B}$ which minimizes $I_m(B) = \|\hat{K}_m(X) - K_m(X;B)\|$. Then

$$I_m(B_m) \le I_m(B_0) \le \|\hat{K}_m(X) - K_{\pi}(X;B_0)\| + \|K_{\pi}(X;B_0) - K_m(X;B_0)\| \qquad (3.7.1)$$

and

$$\|K_{\pi}(X;B_0) - K_m(X;B_m)\| \le \|K_{\pi}(X;B_0) - \hat{K}_m(X)\| + I_m(B_m). \qquad (3.7.2)$$

Assuming identifiability of the class of mixtures generated by the family of component distribution with B and π as parameters, the

strong consistency of B_m follows from Theorem 3.3.2 and the fact that $K_m(X;B_0) \xrightarrow{m} K_\pi(X)$ uniformly. Analogous estimators can be established using the method of moments and the maximum likelihood method.

The final modification takes advantage of the fact that, in general, the random variables observed at adjacent points in time are not independent and, hence, information can be obtained about the unknown parameter by cross correlating successive observations. Raviv [R-2] used this technique to estimate the transition matrix P. His basic tool was a modification of Theorem 3.3.1. Namely, if $g(\cdot,\cdot)$ is an integrable function then

$$P\{\frac{1}{N}\sum_{i=1}^{N} g(X_i,X_{i+1}) \to E_\pi g(X_k,X_{k+1})\} = 1 \qquad (3.7.3)$$

where E_π denotes the expectation when $\underline{p}_1 = \underline{\pi}$. The approach here is based on the fact that

$$K_N(X_N,X_{N+1}) = \sum_{i,j} F_i(X_N)F_j(X_{N+1})p_{ij}P(\lambda_N=i)$$

and

$$P(\lambda_N=i) \xrightarrow{N} \pi_i$$

That is, the sequence of random variables $\{[X_i,X_{i+1}]\}_1^\infty$ is described by a sequence of mixtures of joint distributions which converges to a fixed mixture

$$K_\pi(X_k,X_{k+1}) = \sum_{i,j} F_i(X_k)F_j(X_{k+1})p_{ij}\pi_i$$

Hence a second order estimation strategy analogous to the first order strategy of this chapter can be developed using observations $\{[X_i,X_{i+1}]\}_1^m$ and all past techinques with (3.7.3) in place of theorem 3.3.1. In

this case, the elements of the transition matrix P appear as parameters in the mixture and hence can be included in the list of unknowns to be estimated.[27]

3.8 ADAPTIVE ESTIMATION AND CLASS ESTIMATION

Estimation problems related to the POMS defined in Sec. 2.6 are considered in this section. The transition matrix is assumed to be block-diagonal; hence the system is in one of L noncommunicating classes of states and each class is assumed to satisfy the assumptions of the POMS defined in Sec. 3.1. The component distributions associated with each class are assumed to be unknown to within a parameter as in Sec. 3.1 and the active class of states is unknown also. The problems are to determine which class is active and estimate the unknown parameters of that class.

To formulate the problems more clearly some notation is defined; $\underline{\pi}^i$ is the stationary probability vector corresponding to the i^{th} class;[28] $H_{\pi i} = \{K_{\pi i}(X;B)\}_{B\in\mathcal{B}^i}$ is the set of limit mixtures induced by the family of component cdf's for the i^{th} class with $\mathcal{B}^i$ the corresponding parameter space;[29] i_0 is the value of the index for the active class; B_0 is the true value of the parameter defining the component densities of class i_0. Both B_0 and i_0 are unknown and are to be estimated.

The first problem considered will be that of finding a strongly-consistent estimator for B_0. One strategy is to assume the system is in a particular class and construct an estimate accordingly using the methods of Sec. 3.4 and 3.5. If this is done for all L classes the

---

[27] If P is to be estimated π must be known or a consistent estimator for π must be available. Such an estimator can be obtained from the first order strategy discussed in modification one above.

[28] Each class is being treated as a separate Markov chain.

[29] For simplicity the parameter space is assumed to be the same dimension for all classes. But all arguments apply directly to the more general case.

result in a set of L estimators containing at least one consistent estimator. This procedure is illustrated below using the minimum distance principle.

If $Q_m^i(B) = \|\hat{K}_m(X) - K_{\pi i}(X;B)\|$ then B_m^i, the estimate assuming class i is active, is defined as the element in $\mathcal{B}^i$ which minimizes $Q_m^i(B)$. Since there exists at least one value of i, say $\hat{i}$, such that $\hat{K}_m(X) \overset{m}{\to} K_{\pi}^{\hat{i}}(X;B_0)$ wpl, then, as in Theorem (3.5.1), if $H_{\pi\hat{i}}$ is identifiable $B_m^{\hat{i}} \overset{m}{\to} B_0$ wpl. Therefore a sufficient condition for the adaptive estimator given by (2.6.6) to converge to B_0 is that $H_{\pi i}$ be an identifiable set of mixtures for $i = 1,2,\ldots,L$.

It is important to realize that $\hat{i}$ as defined by

$$Q_m^{\hat{i}}(B_m^{\hat{i}}) \overset{m}{\to} 0 \quad \text{wpl} \tag{3.8.1}$$

may not be unique. Hence B_0 can be estimated even when it is not possible to determine which class is active. The question of sufficient conditions for determining i_0 will now be considered and a more explicit estimate for B_0 will be constructed.

Definition 3.8.1. The set of mixtures $H_{\pi} = \overset{L}{\underset{1}{U}} H_{\pi i}$ is said to be class identifiable if, for any $B_1 \in \mathcal{B}^i$ and $B_2 \in \mathcal{B}^j$

$$K_{\pi i}(X;B_1) = K_{\pi j}(X;B_2) \ \forall X \tag{3.8.2}$$

implies $i = j$.

The above definition can be interpreted as saying that the space of mixtures H_{π} is class identifiable if and only if $\{H_{\pi i}\}_1^L$ is a collection of disjoint sets. Sufficient conditions for class identifiability are given in the Appendix. As indicated by Theorem 3.8.1, class identifiability is the key condition for determining i_0.

THEOREM 3.8.1. If $H_\pi = \bigcup_{i=1}^{L} H_i$ is class identifiable then $\hat{i}_m$ defined by

$$Q(B^{i_m}) = \min_i Q_m^i(B_m^i) \tag{3.8.3}$$

is a strongly-consistent estimator for i_0. Furthermore, if, for each i, $H_{\pi i}$ is identifiable then $B_m^{\hat{i}_m}$ defined by (3.8.3) is a strongly-consistent estimator for B_0.

Proof. By definition and Theorem 3.3.2

$$Q(B^{i_m}) \le Q_m^{i_0}(B_0) \xrightarrow{m} 0 \quad \text{wp1} \tag{3.8.4}$$

Then the subadditivity of the supnorm implies

$$K_{\pi i_m} \xrightarrow{m} K_{\pi i_0} \in H_{\pi i_0} \quad \text{wp1} \tag{3.8.5}$$

where $K_{\pi i_m} = K_{\pi i_m}(X;B^{i_m})$ and $K_{\pi i_0} = K_{\pi i_0}(X;B_0)$. Since for each i the mapping from $\mathcal{B}^i$ onto $H_{\pi i}$, defined by $K_{\pi i}(X;B^i)$, is continuous and $\mathcal{B}^i$ is compact, $H_{\pi i}$ is compact also [R-5]. By the class identifiability of H_π, $\{H_{\pi i}\}_1^L$ is a collection of disjoint sets. Hence, for almost every sequence of observations, there exists an m_0 such that for $m > m_0$ $K_{\pi i_m} \in H_{\pi i_0}$. Otherwise there would exist a subsequence of $\{K_{\pi i_m}\}_1^\infty$ contained in $H_{\pi i_1}$ for some $i_1 \neq i_0$ and converging to $K_{\pi i_0}$ a point not in $H_{\pi i_1}$. This contradicts the compactness of $H_{\pi i_1}$. Thus for $m > m_0$, $i_m = i_0$. If in addition $H_{\pi i_0}$ is identifiable then $\{B^{i_m}\}_{m_0}^\infty$ converges to B_0 wp1, as in Theorem (3.5.1). It follows from both identifiability conditions that for $m \le m_0$ B^{i_m} is well defined.

As indicated in the Appendix identifiability of $H_{\pi i}$ is not a necessary condition for class identifiability of H_π hence i_0 can

be estimated even when B_0 cannot.

3.9 CONCLUSIONS

This chapter has dealt mainly with the problem of finding strongly-consistent estimators for the unknowns in the component distributions of a POMS. The problem was defined in Sec. 3.1 as one of estimating the parameter set B_0 that defines a sequence of mixtures using dependent samples from successive elements in this sequence. The study was restricted to a class of systems with state activity described by a regular Markov chain. As shown in Sec. 3.2, the corresponding sequence of mixtures approaches a limit mixture $K_\pi(X;B_0)$ and the observation process is asymptotically ergodic. These properties were used in Sec. 3.3 to establish tools (extensions of the Law of Large Numbers and the Glivenko-Cantelli Theorem) for estimating $K_\pi(X;B_0)$ and any of its expectations from available observations, thus reducing the estimation problem to the resolution of the limit mixture. This estimation strategy was illustrated with the method of moments and the maximum likelihood method in Sec. 3.4 and the minimum distance principle in Sec. 3.5. The result was a variety of conditions under which the parameters could be estimated and, hence, under which the optimal rule of Chapter II adapts. A specific example illustrating some of these methods and the optimal Bayes estimator of Chapter II was presented in Sec. 3.6. Computer simulations indicated the behavior of the estimators to be typical.

Alternate strategies were proposed in Sec. 3.7. It was shown that these strategies would also handle the case in which the transition matrix P is unknown.

In Sec. 3.8, the adaptive estimation problem and class estimation problems of Sec. 2.6 were solved suboptimally. In the process of establishing conditions under which such estimators could be constructed, the concept of class identifiability was introduced with sufficient conditions for this type of identifiability being given in the Appendix.

Finally, the basic aim of the Chapter has been to put forth a general estimation strategy and illustrate it with examples. It should be clear that many methods not mentioned here [C-4][H-2][S-2][S-3], including nonparametric ones, apply equally well to this problem.

CHAPTER IV

EXAMPLES OF PARTIALLY OBSERVABLE MARKOV SYSTEMS

Examples of Partially Observable Markov Systems (POMS) can be found in the fields of Pattern Recognition and Communication Theory. When the model defined in Sec. 1.1 can be associated with systems in these fields, the decision rules of Chapter II and the estimation schemes of Chapter III lead to a class of decision devices with a learning capability.

This chapter deals mainly with the design of optimum, adaptive signal detectors for a variety of communication systems with unknown signals. The basic approach is to propose a communication system, make the correspondence between it and a POMS with unknown parameter in the component densities, identify the optimal detector, and check critical assumptions that ensure adaption.

The main assumptions that guarantee the existence of an iterative optimal rule which adapts are listed below from Sec. 2.1, 2.5, 3.1 and 3.5.

1. The observation process is state-parameter-conditionally independent.
2. The underlying Markov chain is regular.
3. The Corresponding set of mixtures is identifiable.

Through most of the chapter the transition matrix P will be given and the component densities will be Gaussian with unknown mean. Consequently, in verifying the above assumptions, the following information will be useful.

A regular Markov chain can be characterized by its transition matrix in either of the following two ways [K-3][K-4].

1. There exists a finite integer N, such that P^N has all positive entries, denoted by $P^N > 0$.
2. All but one eigenvalue of P lies inside the unit circle.[30]

As indicated in the Appendix, all the sets of finite mixtures of Gaussian distributions with distinct means are identifiable if constraints are imposed to rule out any ambiguities which may arise in the parameter space. In this chapter it is assumed that, for each example, a parameter prior density which reflects such constraints is available.

The purpose of this chapter is to display the versatility of the model for a POMS and not to investigate each application in detail. In order to display the main ideas clearly, special cases are treated which can be easily generalized. Additional background concerning each problem can be found in the references cited in the corresponding sections.

In Sec. 4.1, a Pattern Recognition System with Markov-dependent pattern activity is introduced and discussed briefly. In Sec. 4.2, a basic communication system with a Markov encoder, memoryless channel, and known synchronization is considered. The assumptions of the basic system are weakened in Sec. 4.3 and 4.4; systems with unknown synchronization and channels with memory are considered. Sections 4.5 and 4.6 deal with variations of the basic system in which synchronization is undefined; namely, systems in which signals arrive at random times. The results of the chapter are discussed generally in Sec. 4.7 and 4.8.

4.1 PATTERN RECOGNITION WITH MARKOV-DEPENDENT PATTERN ACTIVITY

Before attacking any communication systems it is convenient to investigate a more general class of systems and illustrate the format

---

[30] Since the elements in each row of P sum to one, 1 is an eigenvalue.

for utilizing the results of Chapters II and III. In this section, a pattern recognition problem in which pattern activity at one time depends on pattern activity at other times in a Markov manner is shown to be a problem in decision making with a POMS.

The system under study is depicted in Fig. 4.1.1.

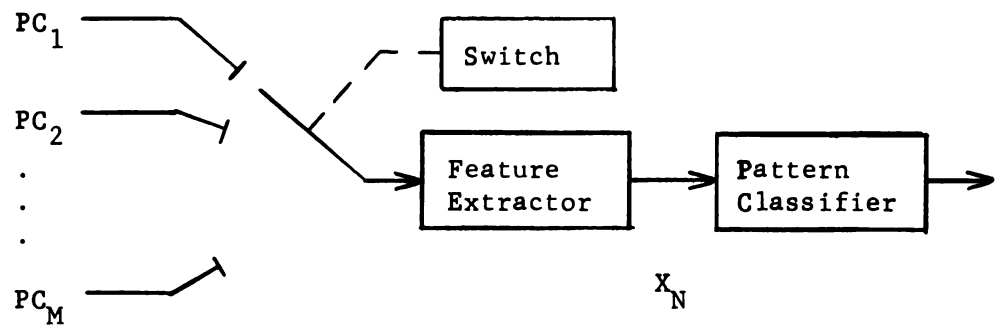

Fig. 4.1.1 A Pattern Recognition System

There are M pattern classes $\{PC_i\}_1^M$. At time N, a sample pattern is randomly chosen according to the probability vector

$$\underline{P}_N = [P(W_1^N),\ldots,P(W_M^N)] \tag{4.1.1}$$

where $P(W_i^N)$ is the probability that PC_i is active at time N and $P(W_i^N|W_j^{N-1},\ldots,W_k^1) = P(W_i^N|W_j^{N-1}) = P_{ji}$ $\forall$ N. The sample pattern is mapped to a point X_N in a finite-dimensional, Euclidean vector space via the feature extractor. Associated with each pattern class is a density $f(X|W_i)$ for X_N when PC_i is active. The pattern classifier makes a decision as to which pattern class generated X_N. Raviv [R-3] applied

such a model to the recognition of characters in text. He assumed all quantities were known or obtainable through supervised learning.

The problem here is to design an adaptive classifier which makes decisions on the basis of observations $X^N = [X_1,\ldots,X_N]$ when $\underline{p}_1$ and $P = [p_{ij}]$ are given but the densities $\{f(X|W_i)\}_1^M$ are unknown. The relation between this problem and that of decision making for a POMS is established by introducing the random variable λ_N which maps the event W_i^N to the interger i. Then $\{\lambda_N\}_1^\infty$ is a first-order, homogeneous Markov chain. The events $\{W_i^N\}_1^M$ have been put into a one-to-one correspondence with the states of a POMS whose state activity is summarized by $\underline{p}_1$ and $P = [p_{ij}]$ and whose observations are governed by the component densities $\{f(X|W_i)\}_1^M$. The problem of classifying feature vectors is that of making decision about the states of a POMS. A variety of adaptive classifiers follow from the decision rules of Chapters II and III.

For example, when the component densities are specified to within a parameter vector B, the minimum probability of error rule as given in Sec. 2.2 is:

$$\text{decide } PC_i \text{ is active at time } N \text{ if}$$

$$P(W_i^N|X^N) \geq P(W_j^N|X^N) \qquad \forall\, j \neq i \tag{4.1.2}$$

where

$$P(W_1^M|X^N) = \int P(W_i^N|X^N,B)\,p(B|X^N)\,dB \tag{4.1.3}$$

As in Sec. 2.3, the posterior probabilities $\{P(W_i^N|X^N)\}_1^M$ can be generated iteratively under Assumption 1 of this chapter. Furthermore, if Assumption 2 and 3 hold

$$P(B|X^N) \xrightarrow{N} \delta(B-B_0) \qquad \text{wp1} \tag{4.1.4}$$

where B_0 is the true parameter value. Hence B_0 is learned and the rule adapts.

It is clear that the estimation schemes and suboptimum rules defined in Chapter III apply as well.

In the remainder of the chapter more explicit examples are given in which the assumptions guaranteeing existence and adaption of optimal rules can be checked.

4.2 ADAPTIVE SIGNAL DETECTION WITH A MARKOV ENCODER

This section considers a particular example of the system treated in Sec. 4.1. A communication system, wherein the signal sent in one time interval depends on that sent during another time interval in a Markov manner, is investigated. Figure 4.2.1 illustrates the system under study.

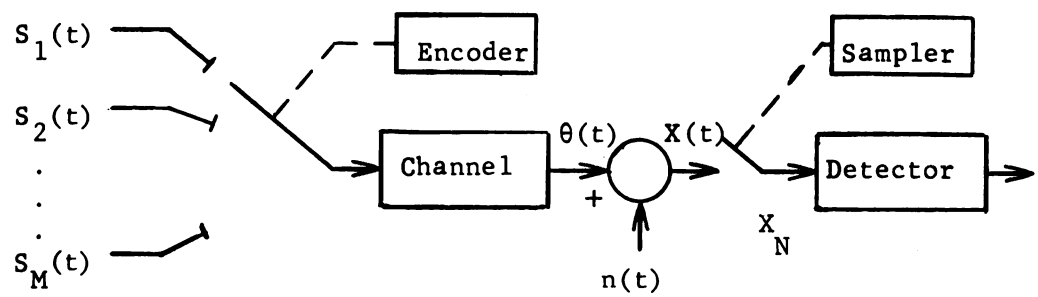

Fig. 4.2.1 A Communication System

Every T seconds a signal is randomly chosen from the set $\{S_i(t);\ 0 \le t \le T\}_1^M$ and sent through a channel which changes it in some fixed but unknown manner. The channel output $\theta(t)$ is corrupted by additive noise and the result is sampled. The observable coordinates are a sequence of time samples $\{X(t_i)\}_1^\infty$ where $X(t_i) = \theta(t_i) + n(t_i)$. The detector uses these observations to determine which signal was sent in a given time interval.

The Basic assumptions are the following

1. The operation of the encoder is described by a matrix of positive probabilities $P = [p_{ij}] = [P(W_j^N/W_i^{N-1})]$ for all N, where W_j^N is the event that the i^{th} signal was sent over the interval $[(N-1)T, NT]$.
2. A set of prior probabilities $\underline{p}_1 = [P(W_1^1),\ldots,P(W_M^1)]$ governing transmission in the first interval is given.
3. The channel is memoryless. Hence, there is no intersymbol interference and the channel output during the N^{th} interval is caused only by the input during that interval. The response to $S_i(t)$ is $\theta_i(t)$.
4. The Noise process is white and Gaussian with zero mean and finite variance.
5. The encoder and detector are synchronized so the time reference is the same for both.
6. Every T seconds a block of ℓ samples is taken at the receiver in accordance with standard sampling theorems; $\underline{X}_N$ is the block taken during the N^{th} interval and $\theta_i(t)$ is characterized by an ℓ-dimensional vector of samples $\underline{\theta}_i = [\theta_{i1},\ldots,\theta_{i\ell}]$.

The correspondence between this system and that of Sec. 4.1 is immediate. The encoder acts as the switching device and the sampler as the feature extractor. When the event W_i^N occurs, the resulting observation is $\underline{X}_N = \underline{\theta}_i + \underline{n}_N$ where $\underline{n}_N$ is a vector of noise samples. Hence, the component densities are Gaussian with means $\{\underline{\theta}_i\}_1^M$. The problem is to establish a detector which makes decisions on the basis of the observations $X^N = [X_1,\ldots,X_N]$ while learning the signal vector $\underline{\theta} = [\underline{\theta}_1,\underline{\theta}_2,\ldots,\underline{\theta}_M]$.

As in Sec. 4.1 the optimum decision rule is:

decide the i^{th} signal was sent in the N^{th} interval if

$$P(W_i^N/X^N) \geq P(W_j^N/X^N) \qquad \forall\, j \neq i \tag{4.2.1}$$

This rule has the learning property indicated by (4.1.4) with B replaced by θ. Only the assumptions remain to be checked. The state-parameter-conditional independence follows from the white Gaussian Noise assumption. This will be the case through the remainder of the Chapter. Hence only Assumptions 2 and 3 will be discussed forthwith. Since $p_{ij} > 0$, the underlying Markov chain is regular. The family of component densities are Gaussian. Hence if the vectors in the set $\{\underline{\theta}_i\}_1^M$ are distinct, the identifiability assumption is satisfied and the rule adapts.

4.3 ADAPTIVE DETECTION WITH INTERSYMBOL INTERFERENCE

When the channel of the system considered in Sec. 4.2 has memory, the signal transmitted in one time interval spills over into other time intervals. Consequently, the received signal at any time is affected by what was sent before. This dependence, under suitable

assumptions, can be shown to be Markovian. Chang [H-1] considered adaptive detection for the binary case in which one of two antipodal signals were sent independently from one time interval to the next. In this section, a more general class of signals transmitted with Markov dependencies is considered.

The system under study is the basic communication system of Fig. 4.2.1 with the channel described by a causal linear filter with impulse response $h(\cdot)$. For simplicity, interference is assumed to be limited to the immediately succeeding time interval, which is expressed as

$$h(t-\tau) = 0 \quad \begin{cases} t > T+\tau \\ t < \tau \end{cases} \tag{4.3.1}$$

The channel input-output relation is given by

$$\begin{aligned} \int_0^T h(t-\tau)S_i(t)\,d\tau &= \theta_i^1(t) \quad 0 \le t \le T \\ &= \theta_i^2(t) \quad T \le t \le 2T \end{aligned} \tag{4.3.2}$$

Let $P(W_{ij}^N)$ be the probability that the i^{th} signal was sent during the N^{th} interval and the j^{th} signal, during the preceeding interval. From the superposition property of linear filters the output of the channel when the event W_{ij}^N occurs is

$$\theta^{ij}(t) = \theta_i^1(NT-t) + \theta_j^2((N+1)T-t) \qquad (N-1)T \le t \le NT$$

and $\underline{X}_N = \underline{\theta}^{ij} + \underline{n}_N$ where $\underline{\theta}^{ij}$ is the vector representation of $\theta^{ij}(t)$, $0 \le t \le T$. Then, using the transition probabilities of Sec. 4.2, the event $\{W_{ij}^N\}$ can be put into one-to-one correspondence with the states of a POMS with Gaussian component densities with means $\{\underline{\theta}^{ij}\}$.

For example if $M = 2$

$$P(W_{11}^N) = P(W_1^N \cap W_1^{N-1}) = p_{11}\, P(W_1^{N-1})$$

$$= p_{11}[P(W_{11}^{N-1}) \cap W_1^{N-2}) + P(W_1^{N-1} \cap W_2^{N-2})]$$

$$= p_{11}\, P(W_{11}^{N-1}) + p_{11}\, P(W_{12}^{N-1})$$

where as defined in Sec. 4.2 W_i^N is the event that the i^{th} signal was sent in the N^{th} interval. This procedure can be repeated for the remaining events in the set $\{W_{ij}^N\}$. Then the probability vector defined by

$$\underline{p}_N = [P(W_{11}^N)P(W_{12}^N)P(W_{21}^N)P(W_{22}^N)]$$

satisfies the iterative relation $\underline{p}_N = \underline{p}_{N-1}P_1$ where

$$P_1 = \begin{bmatrix} p_{11} & 0 & p_{12} & 0 \\ p_{11} & 0 & p_{12} & 0 \\ 0 & p_{21} & 0 & p_{22} \\ 0 & p_{21} & 0 & p_{22} \end{bmatrix}$$

The initial probability vector $\underline{p}_2$ can be calculated from that of Sec. 4.2; e.g. $P(W_{21}^2) = p_{12}\, P(W_1^1)$. When the underlying chain is in the state corresponding to W_{ij}^N the density of the observations is Gaussian with mean $\underline{\theta}_{ij}$.

The optimum decision rule is given by:

decide S_i was sent during the N^{th} interval if

$$P(W_i/X^N) \geq P(W_k/X^N) \qquad \forall\ k \neq i \tag{4.3.4}$$

where

$$P(W_i/X^N) = \sum_j P(W_{ij}/X^N) \tag{4.3.5}$$

and $P(W_{ij}^N/X^N)$ can be generated iteratively as in Sec. 2.3.

The assumptions ensuring adaption are now investigated. Since $p_{ij} > 0$ $i,j = 1,2$, $P_1^2 > 0$ and, thus P_1 is regular. It is clear that, for general M, P_1 retains a structure such that $P_1^2 > 0$. Again, because the component densities are Gaussian, identifiability is insured if the vectors in the set $\{\underline{\theta}^{ij}\}$ are distinct.

When $p_{11} = p_{21} = q_1$ and $p_{22} = p_{12} = q_2$ (this corresponds to the case Chang treated) P^2 is uniform in the columns. Hence $\{\lambda_{2k}\}_1^\infty$ and $\{\lambda_{2k+1}\}_1^\infty$ are independent subsequence of the underlying chains and the corresponding observations are independent. Chang used this fact, which can be arrived at by direct consideration of the model, to construct estimators for the unknown signals. He used the method of moments and his techniques are a special case of those discussed in Sec. 2.4 and 2.7,[31] where if $M = 2$, the unnormalized stationary probability vector is

$$\underline{\pi} = [p_{11}p_{21},\ p_{12}p_{21},\ p_{12}p_{21},\ p_{22}p_{12}]. \tag{4.3.6}$$

4.4 ADAPTIVE DETECTION WITH UNKNOWN SYNCHRONIZATION

When synchronization is unknown in the basic communication system of Sec. 4.2, each block of samples may contain the effects of two signals. Hence there are dependencies between the signal received in one time interval and that in adjacent intervals. Under appropriate assumptions , these dependencies can be shown to be Markovian. Stewart

---

[31] The Third Modification.

[H-3] found the optimum decision rule for determining the true synchronization time when the signals were transmitted independently from one time interval to the next. In this section transmitted signals are Markov-dependent and the optimum decision rule for determining synchronization is shown to follow easily from the results of Sec. 2.6 and 3.8.

The system under study is the basic communication system of Sec. 4.2 without the synchronization assumption. That is, in each block of samples the time sample at which the effect of one signal ends and that of the following signal begins is unknown. Additional assumptions are needed and these will be considered in force for the remainder of the chapter.

1. Time zero is the time the receiver is turned on.
2. The initial probability state vector of Sec. 4.2 governing the transmission of the first signal is the stationary probability vector $\underline{\pi}$.

Assumption 2 implies the time the transmitter is turned on is irrelevant provided it occurs before the receiver is turned on.

If each block is assumed to consist of ℓ samples and the r^{th} sample from the start is the true synchronization time, the N^{th} observation has the form: $\underline{X}_N = \underline{\theta}^r_{ij} + \underline{n}_N$ some i,j where

$$\underline{\theta}^r_{ij} = [\theta_{i,\ell-r+2},\ldots,\theta_{i\ell},\theta_{j1},\ldots,\theta_{j,\ell-r+1}] \tag{4.4.1}$$

is the vector representation of the signal at the output of the channel between $(N-1)T$ and NT. That is, $\underline{\theta}^r_{ij}$ contains the last $r-1$ components of $\underline{\theta}_i$ and the first $\ell-r+1$ components of $\underline{\theta}_j$ where $\underline{\theta}_i$ and $\underline{\theta}_j$ are the vector representations of the channel responses to signals

S_i and S_j respectively.

Let $P(W_{ij}^r \in X_N)$ be the probability that θ_{ij}^r is the channel output in the N^{th} interval and let $P(W^r)$ be the prior probability that r is the true synchronization time. Then, using the transition probabilities of Sec. 4.2 the events $\{W_{ij}^r\}$; $i,j = 1,2,\ldots,M$; $r = 1,2,\ldots,\ell$ can be put into a one-to-one correspondence with the states of a POMS. For example if $M = 2$ and $\ell = 2$

$$P(W_{11}^2 \in X_N) = P(W^2 \cap W_1^k \cap W_1^{k-1})$$

where, as in Sec. 4.2 W_i^k is the event that the i^{th} signal is sent in the k^{th} interval after the transmitter is turned on; k is unknown. Then

$$\begin{aligned} P(W_{11}^2 \in X_N) &= P(W_1^k/W_1^{k-1},W_2)P(W^2 \cap W_1^{k-1}) \\ &= p_{11}[P(W^2 \cap W_1^{k-1} \cap W_1^{k-2}) + P(W^2 \cap W_1^{k-1} \cap W_2^{k-2})] \\ &= p_{11}\,P(W_{11}^2 \in X_{n-1}) + p_{11}\,P(W_{21}^2 \in X_{N-1}) \end{aligned}$$

This procedure can be repeated for all the events $\{W_{ij}^r\}$. Then the probability state vector

$$\underline{P}_N^2 = [P(W_{11}^1)P(W_{22}^1)P(W_{11}^2)P(W_{21}^2)P(W_{12}^2)P(W_{22}^2)]$$

is related to its predecessor by $\underline{P}_N^2 = \underline{P}_{N-1}^2 P_2$ where

$$P_2 = \begin{bmatrix} p_{11} & p_{12} & 0 & 0 & 0 & 0 \\ p_{21} & p_{22} & 0 & 0 & 0 & 0 \\ 0 & 0 & p_{11} & 0 & p_{12} & 0 \\ 0 & 0 & p_{11} & 0 & p_{12} & 0 \\ 0 & 0 & 0 & p_{21} & 0 & p_{22} \\ 0 & 0 & 0 & p_{21} & 0 & p_{22} \end{bmatrix}$$

The initial probability vector $\underline{p}_1$ can be obtained from $\underline{\pi}$ and $\{P(W^r)\}_1^{\ell}$. The corresponding component densities are Gaussian with means $\{\theta_{ij}^r\}$.

The transition matrix is block diagonal with one block associated with each synchronization time. The problem of determining the true synchronization time corresponds to that of class estimation given in Sec. 2.6. The optimum decision rule follows from (2.6.1) and has the form:

decide the true synchronization time is r if

$$P(W^r/X^N) \geq P(W^k/X^N) \qquad \forall\ k \neq r$$

where

$$P(W^r/X^N) = \sum_{i,j} P(W_{ij}^r/X^N)$$

Conditions under which

$$P(W^r/X^N) \xrightarrow{N} \Delta_{r_1 r_0}, \qquad \text{wpl}$$

where r_0 is the true synchronization time, are now considered. In general, P_2 will have ℓ-1 blocks exactly the same (those corresponding to synchronization time $r = 2,\ldots,\ell$). These have the same form as P_1 considered in Sec. 4.3 and hence are regular. The remaining block (for $r = 1$) has all positive elements and is regular also. Since for $\ell > 2$ the blocks are not distinct and the component densities belong to the same family, then as indicated in Sec. 3.8 and the Appendix r_0 cannot be determined unless prior information is available concerning the parameters of each class. However, this POMS has some special properties that can be exploited. The observations

from one class are related to those in others(when they are active) through a shifting procedure. Hence, an empirical cdf for the limit mixture of each class can be constructed for minimum distance estimation. This observation was used by Stewart in the i.i.d. case to establish that r_0 could be uniquely determined if $\{\underline{\theta}_{ij}^{r}\}$ contained at least m+1 distinct vectors for each r. Since the identifiability problem is the same for the Markov case, his result can be extended via Theorem 3.3.2.

4.5 M- ary ADAPTIVE DETECTION OF SIGNALS WITH RANDOM ARRIVAL TIMES

In Sec. 4.2-4.4, the periodic behavior of the encoder allowed observations to be processed in blocks of known size, whether or not synchronization was known. In this section and the next, the uncertainty concerning the signal arrival time is greatly increased and the samples must be processed one at a time. The detection of signals of known duration but with random arrival times is considered. For random lengths of time no signal is transmitted and only noise is received.

Stewart [H-3] found the optimum receiver for a case in which signals were transmitted independently in time but was unable to find a suboptimum solution and, therefore, could not prove adaption. In this section, the optimum receiver for the Markov-dependent case is established and convergence follows immediately from previous results.

The system under consideration is essentially the basic communication system of Sec. 4.2 with an additional "no signal" input which can be active for a random length of time. Signals are transmitted

for a duration T and no transitions of the encoder are allowed during this time. A typical signal of the output of the channel is depicted in Fig. 4.5.1.

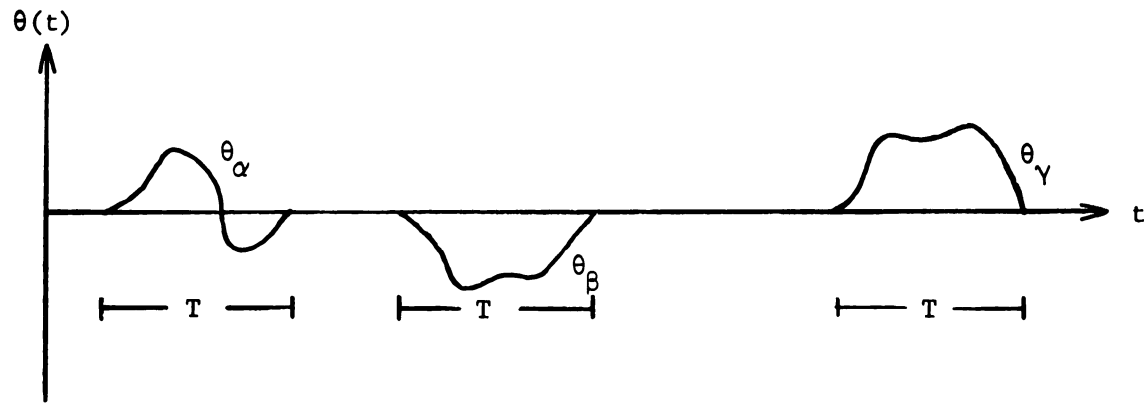

Fig. 4.5.1 Channel Output for an M-ary Signal set with Random Arrival Times

Then $X_k = \theta(t_k) + M(t_k)$ is the sample at time t_k. The sampler is assumed synchronized; i.e., signals can arrive only at sampling instants. Thus, $\theta(t_k)$ is either 0 or θ_{ij}, the j^{th} component of the signal vector characterizing $\theta_i(t)$; $0 \le t \le T$.

Let $P(W_{ij}^k)$ be the probability that θ_{ij} is active in X_k and let $P(W_0^k)$ be the probability that noise alone is present. Then, as in previous sections, the events $\{W_{ij}^k\}$, W_0^k can be associated with the state of a POMS with Gaussian component densities with means $\{\theta_{ij}\}$ and 0.

When $\ell = M = 2$ the probability state vector

$$\underline{P}_k = [P(W_0^k)P(W_{11}^k)P(W_{12}^k)P(W_{21}^k)P(W_{22}^k)]$$

can be generated by $\underline{P}_k = \underline{P}_{k-1}P_3$ where

$$P_3 = \begin{bmatrix} p_{00} & p_{01} & 0 & p_{02} & 0 \\ 0 & 0 & 1 & 0 & 0 \\ p_{10} & p_{11} & 0 & p_{12} & 0 \\ 0 & 0 & 0 & 0 & 1 \\ p_{20} & p_{21} & 0 & p_{22} & 0 \end{bmatrix}$$

The probabilities p_{0i}, and p_{i0} are those governing transitions (when they can occur) between the signal and noise states of the encoder, from one sampling time to the next.

The following posterior probabilities can be used to make optimal decisions as to which signal, if any, is active at time k.

$$P(W_i^k) = \sum_j P(W_{ij}^k / X^k) \qquad i = 0,1,\ldots,M$$

For ease in verifying assumptions, let $p_{ij} = q_j$ $j = 0,1,\ldots,M$ (This is the case considered by Stewart). For this case with $M = \ell = 2$ P_3 has eigenvalues 0, 0, 0, 1, q_0-1 and P_3 is regular. With general M and L, the characteristic polynomial of P_3 can be shown to be

$$g_{m\ell}(s) = (-1)^{m\ell} s^{(m-1)\ell+1}(-s^{\ell} + q_0 s^{\ell-1} + q)$$

where $q = 1 - q_0 = \sum_1^M q_i$. Since $q_{m\ell}(s) = 0$ has only one root on the unit circle[32], s = 1, P_3 is regular. In the most general case P_3 has the same structure, with respect to non zero entries, as in the previous case. Hence the type of state activity is the same and P_3 is regular.[33]

Identifiability conditions follow as before. If $\{\theta_{ij}\}$ is a set of distinct elements, all unknown parameters can be determined.

---

[32]The only value of a which satisfies $q_{mL}(e^{ja}) = 0$ is a = 0.

[33]See Sec. 4.7.

4.6 ADAPTIVE DETECTION OF SIGNALS WITH RANDOM ARRIVAL TIMES

This section treats a variation of the problem defined in Sec. 4.5. Instead of employing a randomly-chosen sequence of signals with random spacing, one signal is randomly chosen and transmitted repeatedly at random times. The problem is to design detectors to determine when the signal is present and which signal is being sent. Nolte [N-1] found a detector for the case in which all the signals are known. However, he did not discuss the conditions for adaption; i.e., for convergence to the detector that would be used if the identity of the signal being sent were known. Here, the signals are unknown but other prior information is assumed available to make the problem meaningful.

The system under study is basically the same as that in Sec. 4.5 except that here the encoder switches between a fixed signal and noise. A typical signal at the channel output is shown in Fig. 4.6.1.

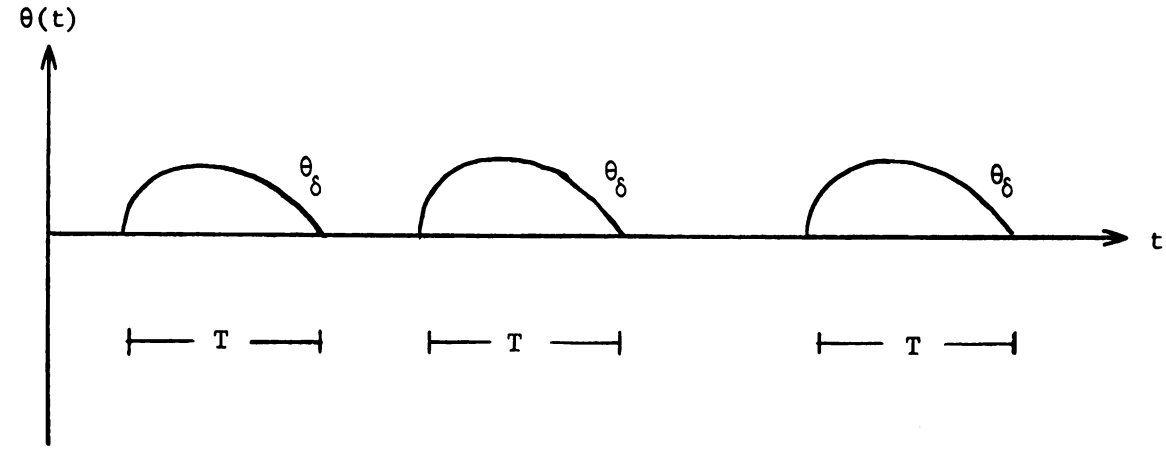

Fig. 4.6.1 Channel Output for a Signal with Random Arrival Times

When the i^{th} signal is being sent $X(t_k) = X_k = \theta(t_k) + n(t_k)$ where $\theta(t_k)$ is either 0 or θ_{ij} for some j; θ_{ij} is the j^{th} component of the signal vector corresponding to the i^{th} signal $\theta_i(t)$. Let $P(W_i)$ be the prior probability that signal i is the one being sent repeatedly.

As in Sec. 4.5, $P(W_{ij}^k)$ is the probability that $X_k = \theta_{ij} + n(t_k)$ and $P(W_{i0}^k)$ is the probability that $X_k = n(t_k)$ and signal i is the one being sent. Then, the event $\{W_{ij}^k\}$ can be associated with states of a POMS with a block diagonal transition matrix, each block corresponding to the event that a particular signal is being sent repeatedly.

For example, when $M = \ell = 2$ the probability state vector is

$$\underline{P}_N = [P(W_{10}^k), P(W_{11}^k), P(W_{12}^k), P(W_{20}^k), P(W_{21}^k), P(W_{22}^k)] \tag{4.6.1}$$

and the transition matrix is

$$P_4 = \begin{bmatrix} \nu_1 & 1-\nu_1 & 0 & & & \\ 0 & 0 & 1 & & & \\ \nu_1 & 1-\nu_1 & 0 & & & \\ & & & \nu_2 & 1-\nu_2 & 0 \\ & & & 0 & 0 & 1 \\ & & & \nu_2 & 1-\nu_2 & 0 \end{bmatrix} \tag{4.6.2}$$

where $\nu_i = P(W_{10}^k / W_{i0}^{N-1}) = P(W_{i0}^k / W_{i2}^k)$ are assumed given. Again, the component densities are Gaussian with means $\{\theta_{ij}\}$ and 0.

Then, as indicated in Sec. 2.6, under appropriate assumptions to be discussed presently

$$P(W_i / X^k) = \sum_j P(W_{ij} / X^k) \to \Delta_{ij_0} \quad \text{wp1} \tag{4.6.3}$$

where j_0 is the index of the true signal being sent and $P(W_i)$ is the probability that i^{th} signal is being sent. The probabilities $\{P(W_i/X^k)\}_1^M$ can be used to make decisions as to which signal is being sent.

Now if the i^{th} signal is assumed the one being sent the decision rule is

decide s_i present if

$$L(X^k/W_i) = \sum_{j=1}^{\ell} \frac{P(W_{ij}/X^k W_i)}{P(W_{i0}/X^k W_i)} > 1 \tag{4.6.4}$$

Consequently an adaptive detector is given by

decide a signal present if $L(X^k) > 1$

where

$$L(X^k) = \sum_{z=1}^{M} L(X^k/W_i) P(W_i/X^k) \tag{4.6.5}$$

and

$$L(X^k) \rightarrow L(X^\infty/W_{i_0}, \theta_{i_0}) \tag{4.6.6}$$

The rule (4.6.4) is analogous to that given by Nolte but it is not the optimal rule for detecting a signal. The optimal rule uses the likelihood ratio

$$\sum_{i=1}^{M} \sum_{j=1}^{\ell} P(W_{ij}/X^k) \Big/ \sum_{i=1}^{M} P(W_{i_0}/X^k)$$

which has adaption properties similar to (4.6.6).

The conditions for convergence will now be investigated. The i^{th} block of P_4 corresponds to a special case ($M = 1$) of P_3 considered in Sec. 4.5. Hence, P_4 is regular.

In checking identifiability it is important to realize that there are two estimation problems involved. The first is concerned with determining which signal is being sent and involves class identifiability. The second involves the estimation of the signal being sent and regular identifiability. As indicated in Sec. 3.8, (4.6.3) can hold without (4.6.6) being true and vice versa.

For example, let $M = \ell = 2$. Then the stationary probability vector for the class corresponding to the i^{th} signal is proportional to $[\nu_i \ 1-\nu_i \ 1-\nu_i]$. Hence, according to the discussion in the Appendix, (4.6.3) will be true if $\nu_1 \neq \nu_2$ and, for each i, $\{\theta_{ij}\}$ is a set of distinct elements. If $\nu_1 = \nu_2$, additional prior information is needed on the parameters. It must be known that θ_1 and θ_2 lie in disjoint regions of the parameter space; e.g., the signal set might be antipodal. If such information is not available, the signals cannot be distinguished, in general.

On the other hand, if, for each i, $\{\theta_{ij}\}$ is a set of distinct elements, $\underline{\theta}_{i_0}$ can be learned and (4.6.6) will obtain even if the class of the signal is indeterminable. If no special prior information is available concerning the signal from each class, then one might as well design a detector for one unknown signal. The rule of Sec. 4.5 could be used with $M = 1$. The important point is that the rules considered in this section provide a means of using this prior information when it is available.

When θ is known (the case treated by Nolte) it is clear from the Appendix that the existence of the set $\{\underline{\theta}_i\}$ of distinct vectors is sufficient for adaption.

4.7 REMARKS

This section considers briefly some points that apply generally to the contents of this chapter.

In attempting to treat a variety of situations in a uniform manner, simplifying assumptions were made and the models were slightly contrived. Consequently, the results generalize in many respects and are intended to include other situations that give rise to similar decision problems. The emphasis in these applications should be on the received signals with the encoder and channel serving as a convenient way of accounting for the generation of unknown signals in a Markov manner. From this point of view, the results of Sec. 4.3 indicate that the Markov dependencies between the received signals of other sections could be due to intersymbol interferrence; the randomly arriving signals of Sec. 4.5 or 4.6 could, for example, be seismic waves; any signal space representation of the received signal in Sec. 4.2 will serve as well as time samples to make decisions. It is also clear that the Gaussian noise assumptions can be weakened and the number of unknowns can be increased.

With regard to the identifiability conditions, the following observations are important. The conditions stated in each section are sufficient to ensure that all parameters can be learned and effective decisions can be made on all the states of the corresponding POMS. However, in many cases, (Sec. 3.3-3.6) the events of interest consist of a union of other events. To make effective decisions, it is not necessary that all the component densities corresponding to the events in this union be distinguishable. Thus, depending on the inferrence

problem of interest and the available prior information, conditions for adaption can be considerably weakened, with Sec. 3.8 and the Appendix providing the guidelines.

Verifying the regularity of the underlying Markov chains may have appeared to be a formidable task. However, it is well known that the regularity of a Markov chain can generally be determined from the structure of the corresponding transition matrix. While the eigenvalues give useful information concerning the system activity, they do not have to be computed to verify this assumption. Furthermore, the behavior of a general class of systems can usually be summarized by a simple example. This point of view was not developed in this chapter but was implicit in Sec. 4.5.

Finally, while the optimum solutions given in this chapter provide a reference for comparison, they are generally undesirable from the view point of practical engineering. Among the suboptimum solutions suggested by the estimation techniques of Chapter III it appears the method of moments would yield the most fruitful results. The success that Chang and Stewart have had with this method in developing low-memory estimators for cases of practical interest indicate that similar results could be obtained here.

4.8 CONCLUSIONS

The aim of this chapter has been to display the versatility of the model for a POMS and to exhibit how the results of Chapters II and III can be applied. Several communication systems were shown to be POMS. These include systems with

1. A Markov information source
2. A channel with memory
3. unknown synchronization
4. signals with random arrival time
5. combinations of the above

For all these cases the form of the optimal detector for unknown received signal was shown to follow directly from the results of Chapter II and the conditions for adaptions from Chapter III. Although optimal decision making was emphasized in the above examples, it is implied that the estimation schemes and suboptimal decision rules apply as well.

In addition to providing quick solution, the technique of formulating these inference problems as those of decision making for a POMS has other advantages. First, it provides a common model for a variety of seemingly different systems. This facilitates comparison and helps focus analysis efforts in one direction. Results developed for general POMS apply to all of the above systems. Next, it illustrates clearly the nature of the estimation problem involved. Whereas, the properties of the observations are not always clear, in an ad hoc formulation, the mixture approach used in this thesis clearly defines conditions for the existence of estimators and suggests a wealth of techniques. Finally, it brings into play the powerful tools of Markov chain theory. Once a state space and transition matrix have been established, a great deal can be inferred about system state activity.

CHAPTER V

GENERAL CONCLUSIONS

In this chapter, the main contributions of the Thesis are reviewed and suggestions are made for future research.

5.1 REVIEW

This Thesis has been concerned with several inference problems related to a class of Partially Observable Markov Systems. Generally, the results represent an extension of previous work on unsupervised learning and adaption from the i.i.d. case to a particular case with dependent, non-stationary observations. However, additional inference problems not well established for the i.i.d. case were generalized and solved here also; namely those of class estimation and adaptive estimation treated in Sec. 2.6 and 3.8. The basic goal has been to construct estimators and decision rules and to state conditions under which they perform effectively.

In Chapter II, Bayes Decision-Theoretic concepts were used to develop optimal solutions when the component densities are defined by an unknown parameter set. Large sample theory was used in Chapter III to establish suboptimum solutions. The basic estimation problem was one of mixture resolving and a general strategy was developed for extending estimation techniques developed for the i.i.d. case to the more general case studied here. In general the results display many similarities with the i.i.d. case. The dominant role of mixtures and

identifiability, the implementation difficulties of the optimum rule and the need for prior information are all general characteristics of nonsupervisory problems.

Chapter IV showed several communication systems of current interest to be POMS. Consequently adaptive detectors and estimators were easily established along with conditions for effective operation. This unifying approach to solving a previously troublesome class of problems represents a major contribution of the Thesis.

5.2 EXTENSIONS

The similarities between the results obtained here and previous work in both estimation and Markov chain theory suggest certain natural extensions that used to be investigated.

First is a class of inference problems with time varying parameters. For example, as an extension of the optimal i.i.d. case Braverman [B-3] and Fralich [F-1] considered parameter changes summarized by the difference equation of the form

$$B_{M+1} = B_M + \Delta_M$$

where $\{\Delta_M\}$ is a sequence of independent random variables. Their ideas are applicable to the problem treated here [H-5].

Next is the problem of developing estimators that are easier to implement than those given here. Sakrison [S-1] has used stochastic approximation techniques to solve the likelihood equation. The result is a simple iterative low memory estimator. This method applies to ergodic observation processes which suggests that it could be extended to the asymptotically ergodic case treated in this thesis.

Finally, most of the results in this Thesis apply to POMS whose state activity is described by a regular Markov chain. Although as demonstrated in Chapter IV this represents a useful class of systems it would be desirable to extend the results to include ergodic chains. In so far as those properties which effect the proof of Theorem 3.3.1 are concerned, regular and ergodic chains are similar and it appears the general estimation strategy can be extended.

5.3 SOME INTERESTING PROBLEMS

In an attempt to exploit the similarities between the i.i.d. case and that of a general POMS, several interesting problems have been ignored. Most of these problems arise from the fact that, unlike the i.i.d. case, the optimum decision rule for a POMS has a changing structure (as a function of the observations) even when all quantities in the model are known. This causes three main difficulties.

The first problem is concerned with implementation of suboptimum rules. Unless $P(\lambda_N=i/X^N,B)$ is stored as a function of B (This would lead to the same storage problems as the optimum rule) using $P(\lambda_N=i/X^N,B_N(X^N))$ to make decisions at the N^{th} step requires storage of X^N and an iteration over N steps using the schemes of Sec. 1.2. This is the case regardless of what is available from the $(N-1)^{th}$ step. Hence the memory and number of computations grow with N. In an attempt to overcome this problem Raviv [R-2] has shown, for P unknown and $\underline{p}_1 = \underline{\pi}$, that a fixed number of past samples can be used to construct decision rules with an asymptotic risk arbitrarily close to the corresponding risk for known P. The question of how many past

observations are needed in a practical situation remains open.

The second problem is that of proving adaption of both optimal and suboptimal rules. For example one would like to show

$$\left| P(\lambda_N = i/X^N, B_N(X^N)) - P(\lambda_N = i/X^N, B_0) \right| \to 0 \quad \text{wp1}$$

This is a nontrivial statement and again the problem can be traced to the varying structure of the rule. Raviv by using only a fixed number of past observations had a fixed form rule (given P) and adaption followed.

The third problem is that of computing probability of error. Even when allquantities are known and a very simple example is used error calculations are prohibitive [D-3]. This suggests that computer simulation must be used to evaluate adaptive decision making devices.

The above problems indicates that a profitable result would be some practical measure of dependency between the samples. If such a tool were available effective suboptimum rules could be constructed with desirable implementation properties; the look-ahead mode of decision making could be better evaluated; and the effect of initial system behavior on asymptotic decision modes could be determined. At present computer simulation and intuition are the only guides.

BIBLIOGRAPHY

BIBLIOGRAPHY

[A-1] Aoki, M., Optimization of Stochastic Systems, Academic Press, New York, 1967.

[B-1] Billingsley, P., "Statistical Methods in Markov Chains", Ann. Math. Statist., Vol. 32, pp. 12-40, 1961.

[B-2] Blischke, W.R., "Moment Estimators for the Parameters of Two Binomial Distributions," Ann. Math. Statist., Vol. 33, pp. 444-454, 1962.

[B-3] Braverman, D., Machine Learning and Automatic Pattern Recognition, Stanford Electronics Laboratories, Technical Report No. 2003-1, Feb., 1961.

[C-1] Chang, R.W. and J.C. Hancock, "On Receiver Structures for Channels Having Memory," IEEE Trans. on Information Theory, Vol. IT-12, No. 4, pp. 463-468, Oct., 1966.

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

[C-3] Choi, K., Estimates for the parameters of a finite mixture of distributions, University of Missouri, Technical Report No. 18, 1966.

[C-4] Cooper, D.B., On the Existence of Nonsupervised Adaptive Signal Detectors and Detector Estimation using Stochastic Approximation Methods, Ph.D. Dissertation, Columbia University, April 1966.

[C-5] Cooper, D.B., and P.W. Cooper, "Nonsupervisory Adaptive Signal Detection," Information and Control, Vol. 7, No. 3, pp. 416-444, Sept., 1964.

[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] Doob, J.L., Stochastic Processes, Wiley, New York, 1953.

[D-3] Drake, A.W., "Observation of a Markov Source Through a Noisy Channel," presented at the IEEE Symposium on Signal Transmission and Processing, Columbia University, May 1965.

[F-1] Fralick, S.C., "The Synthesis of Machines which Learn Without a Teacher," Stanford Technical Report No. 61308-9, April 1964.

[H-1] Hancock, J.C., and R.W. Chang, "Unsupervised Learning Receivers for Binary Channels with Intersymbol Interference," presented at the IEEE Symposium on Signal Transmission and Processing, Columbia University, May 1965.

[H-2] Hancock, J.C., and E.A. Patrick, "Learning Probability Spaces for Classification and Recognition of Patterns with or without Supervision", Purdue University Report TR-EE-65-21, Nov. 1965.

[H-3] Hancock, J.C., and T.L. Stewart, Parameter Estimation with Unknown Symbol Synchronization, Purdue University Report TR-EE-67-1.

[H-4] Hancock, J.C., and P.A. Wintz, Signal Detection Theory, McGraw-Hill, 1966.

[H-5] Hilborn, C.G., and D.G. Lainiotis, "Optimal Unsupervised Learning Multicategory Dependent Hypothesis Pattern Recognition," IEEE Trans. on Information Theory, Vol. IT-14, No. 3, pp. 468-470, May 1968.

[K-1] Kakalik, J.S., Optimum Policies for Partially Observable Markov Systems, Technical Report No. 18, Operations Research Center, Massachusetts Institute of Technology, Oct. 1965.

[K-2] Kale, B.K., "On the Solution of the Likelihood Equation by Iteration Processes," Biometrika, Vol. 48, pp. 452-456, 1961.

[K-3] Karlin, S., A First Course in Stochastic Processes, Academic Press, New York, 1966.

[K-4] Kemeny, J.G., and J.L. Snell, Finite Markov Chains, Van Nostrand, 1960.

[L-1] Loeve, M., Probability Theory, Van Nostrand, New York, 1955.

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

[M-2] Mendel, J.M., "A Survey of Learning Control Systems," I.S.A. Transactions, pp. 297-303, July 1966.

[N-1] Nolte, L.W., "An Adaptive Realization of the Optimum Receiver for a Sporadically Recurrent Wave Form in Noise," IEEE Trans. on Information Theory, Vol. IT-13, pp. 308-401, April 1967.

[P-1] Patrick, E.A., and J.C. Hancock, "Nonsupervised Sequential Classification and Recognition of Patterns," IEEE Trans. on Information Theory, Vol. IT-12, No. 3, pp. 362-372, Oct. 1966.

[P-2] Patrick, E.A., "On a Class of Unsupervised Estimation Problems," IEEE Trans. on Information Theory, Vol. IT-14, No. 3, pp. 407-415, May 1968.

[P-3] Pearson, K.P., "Contributions to the Mathematical Theory of Evolution," Phil. Trans. Roy. Soc., 185A 71-110, 1894.

[R-1] Rao, C.R., Advanced Statistical Studies in Biometric Research, John Wiley and Sons, New York, 1952.

[R-2] Raviv, J., "Decision Making in Incompletely Known Stochastic Systems," Int. J. Eng. Sci., Vol. 3, pp. 119-140, Pergamon Press, 1965.

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

[R-4] Rider, P.R., The Method of Moments Applied to a Mixture of Two Exponential Distributions, Ann. Math. Statist., Vol. 32, pp. 143-147, 1961.

[R-5] Royden, H.L., Real Analysis, Macmillan, New York, 1963.

[S-1] Sakrison, D.J., "Stochastic Approximation: A Recursive Method for Solving Regression Problems," Advances in Communication Systems, Vol. 2, 1966, Academic Press, New York, 1966.

[S-2] Sammon, J.W., "An Adaptive Technique for Multiple Signal Detection and Identification," IEEE Convention Record, 1967.

[S-3] Scudder, H.J., "Adaptive Communication Receivers," IEEE Trans. on Information Theory, Vol. IT-11, pp. 167-174, April 1965.

[S-4] Slansky, J., "Learning Systems for Automatic Control," IEEE Trans. on Automatic Control, Vol. AC-11, No. 1, pp. 6-19, Jan. 1966.

[S-5] Spragins, J.D., Reproducing Distributions for Machine Learning, Stanford Electronic Lab., Technical Report No. 6103-7, Nov. 1963.

[S-6] Spragins, J.D., "Learning Without a Teacher," IEEE Trans. on Information Theory, Vol. IT-12, No. 2, pp. 223-230, April 1966.

[T-1] Teicher, H., "On Mixtures of Distributions," Ann. Math. Statist., Vol. 31, pp. 55-73, 1961.

[T-2] Tucker, H.G., A Graduate Course in Probability Theory, Academic Press, 1967.

[W-1] Wilks, S.S., Mathematical Statistics, John Wiley and Sons, New York, 1962.

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

APPENDIX

APPENDIX

Identifiability and Prior Information for Unsupervised Learning

In Sec. 3.5 and 3.8, it was shown that when the set of mixtures which arise in the estimation problem of Sec. 3.1 is identifiable, the unknown parameter vector B_0 can be uniquely determined from the observations. To ensure this condition constraints must be imposed on the family of component densities and the parameter space. Thus in a particular decision making problem a certain amount of prior information is needed to guarantee solution. This Appendix deals with sufficient conditions for identifiability. Most of the results are taken from the literature but are presented from a slightly different point of view, more suitable for the problems of interest in this Thesis. Sufficient conditions are also established for class identifiability introduced in Sec. 3.8.

The Uniqueness of Representation Property for Finite Mixtures

Let $\mathfrak{F} = \{F(X;\alpha)\}_{\alpha \in A}$ be a family of joint densities indexed by a point α taking values in a subset of a finite dimensional Euclidean vector space A. Let

$$H^k = \{H(X) = \sum_{i=1}^{k} C_i F(X;\alpha_i),\ C_i > 0,\ \sum_{1}^{k} C_i = 1,\ F(X;\alpha_i) \in \mathfrak{F}\}$$

where the $\{\alpha_i\}_1^k$ is a distinct set of elements. Then $\mathfrak{H} = \bigcup_{1}^{\infty} H^k$ is the set of all finite mixtures of the family $\mathfrak{F}$. The set $\mathfrak{H}$ is said

to have the uniqueness of representation property (urp) if

$$\sum_{i=1}^{\hat{k}} \hat{C}_2 F(X;\hat{\alpha}_i) = \sum_{i=1}^{k} C_i F(X;\alpha_i) \tag{A-1}$$

implies $\hat{k} = k$ and for each i, $1 \le i \le k$ there is some j, $1 \le j \le k$ such that $C_i = \hat{C}_j$ and $\alpha_i = \hat{\alpha}_j$. The URP can be restated as saying there is a one-to-one correspondence between each set of allowable points $\{C_i,\alpha_i\}_1^k$ and the mixture they generate. As indicated by the following theorem this property can be characterized by $\mathfrak{F}$.

Theorem A-1. A necessary and sufficient condition that $\mathfrak{A}$ have the URP is that $\mathfrak{F}$ be a linearly independent set over the field of real numbers.

This theorem can be used to establish the URP for the set of finite mixtures generated by the following families.

1. The family of n-dimensional Gaussian cdf's indexed by the mean and/or the covariance matrix.
2. The family of n-dimensional exponential cdf's indexed by the exponent constant.
3. The translation parameter family induces by any cdf with a bilateral Laplace transform.

That is any finite set of distinct elements in each of the above families is a linearly independent set.

Identifiability and the Estimation Problem

For the estimation problem Sec. 3.1 the class of mixtures of interest is $H_\pi = \{K_\pi(X;B)\}_{B\in\mathcal{B}}$ where

$$K_\pi(X;B) = \sum_{i=1}^{M} \pi_i f(X;B_i) \tag{A-2}$$

In (A-2) $B = [B_1,\ldots,B_M]$, $f(X;B_i) \in \mathfrak{F}$, and $\mathcal{B}$ is a subset of M-dimensional Euclidean space containing only vectors with distinct components. Then H_π is said to be an identifiable class of parameter indexed mixtures if the mapping defined by (A-2) say Q_π is a one-to-one mapping from $\mathcal{B}$ onto H_π. Then if $K_\pi(X) \in H_\pi$ there exists a unique vector B_0 such that $K_\pi(X;B_0) = K_\pi(X)$.

Since $H_\pi \subset H$, if H has the URP so does H_π. However the URP guarantees the uniqueness of B_0 only to within an equivalence class. That is permutations of the components of B_0 might result in another solution vector. This would be the case if all the components of π were not distinct. Thus in order to guarantee identifiability as defined here a constraint on the parameter space is needed as well as URP. In a practical situation this constraint will be a reflection of prior information on a particular problem. For example, if $M = 2$, $\pi_1 = \pi_2 = \frac{1}{2}$ and it is known $B_1 > B_2$, where B_i is the parameter associated with state i of the system, then with the URP B_0 can be uniquely determined by minimum distance estimation. It is important to realize also that in order for the Bayes algorithm to learn B_0 these constraints must be reflected in the prior distribution $P_0(B)$. If one is not interested in using the estimates for decision making then constraints can be arbitrarily imposed to get a unique solution vector. The definition of identifiability and all the above remarks can be extended to the case in which π is unknown and the parameter space is 2M-dimensional.[34]

---

[34] Identifiability is usually defined for this larger class of mixtures [H-2] and the above definition is a consistent modification.

Class Identifiability

If $\mathfrak{F}^i$ is the family of component CDF's associated with the i^{th} class then as in Sec. 3.8 $H_{\pi^i} = \{K_{\pi^i}(X;B^i)\}_{B^i \in \mathcal{B}^i}$ is the corresponding set of mixtures with parameter space $\mathcal{B}^i$ and probability state vector π^i. According to definition 3.3.1 $H_\pi = \bigcup_{i=1}^{L} H_{\pi^i}$ is said to be class identifiable if $\{H_{\pi^i}\}_1^L$ are disjoint subsets of H_π. Then from Theorem A-1 and a simple contradiction argument any of the following conditions are sufficient for class identifiability of H_π.

1. $\mathfrak{F} = \bigcup_{i=1}^{L} \mathfrak{F}^i$ is a linearly independent set, $\{\pi^i\}_1^L$ is a set of vectors distinct to within permutations on the components.
2. $\mathfrak{F}$ is a linearly independent set by $\mathcal{B}^i \cap \mathcal{B}^j = \phi \quad i \neq j$
3. $\mathfrak{F}$ is a linearly independent set and $\bigcap_{i=1}^{L} \mathfrak{F}^i = \phi$

Assuming $\mathfrak{F}$ is a linearly independent set, conditions 1 and 3 above indicate that classes of states can be distinguished when the stationary probability vector associated with different classes are distinct or the component densities associated with different classes have distinct forms. However condition 2 indicates that if sufficient prior information is available concerning the parameters associated with each class the classes can be distinguished even when the class transition matrices are equal and the component densities for each class have the same form.

The above conditions are sufficient and it is clear that they can be weakened. For example if $\mathfrak{F}^i$ is not a linearly independent set, its mixtures do not necessarily equal those corresponding to other classes. Also the components of B_0^i need not be distinct to distinguish classes. In fact <u>when π is known</u> the distinctness of

the components of B_0 is needed only to ensure the system states can be distinguished in some sense and is not necessary to estimate B_0. Finally when the class transition matrices are distinct but the corresponding stationary probability vectors are not, a second order estimation strategy (Sec. 3.7) may still lead to an estimator for i_0.