BAYESIAN DECISION MAKING AND LEARNING FOR CONTINUOUS - TIME MARKOV SYSTEMS

Thesis for the Degree of Ph. D.

MICHIGAN STATE UNIVERSITY

ERDAL PANAYIRCI

1970

This is to certify that the

thesis entitled

BAYESIAN DECISION MAKING AND
LEARNING FOR CONTINUOUS-TIME
MARKOV SYSTEMS

presented by

ERDAL PANAYIRCI

has been accepted towards fulfillment
of the requirements for

Ph.D. degree in Electrical Engineering & Systems Science

Major professor

Date November 17, 1970

ABSTRACT

BAYESIAN DECISION MAKING AND LEARNING FOR CONTINUOUS-TIME MARKOV SYSTEMS

By

Erdal Panayırcı

This thesis is concerned with Bayesian decision making and learning algorithms for a particular problem in parametric pattern recognition in which each of a finite set of pattern classes is characterized by a continuous-time, discrete-state Markov process. The basic problem considered is that of determining rules for making decisions about the identity of the active pattern class based upon observation of a sample function in some finite interval. The stationary transition probability matrices for the processes in question are the parameters of the pattern classes.

Statistical decision theory is employed throughout to develop optimal solutions. In the first part of the thesis, Bayes-optimum decision rules are derived under a perfect observation mechanism (noiseless case). The observed quantities are the sojourn times in the states and the state numbers themselves. Using classified samples from each pattern class, an algorithm for supervised learning is presented and the existence of reproducing prior densities for the parameters is demonstrated. Particularly useful results are the formulations of recursive, computationally simple parametric forms for the posterior densities of the unknown parameters and for the optimum decision rules. The simulation of a specific example shows

the empirical probability of error for different amounts of training data and demonstrates the inherent practicality of the results.

The problem of computing probability of error is investigated extensively for the noiseless case. The exact probability of error, as well as lower and upper bounds and asymptotic expressions, are established for several cases. Conditional error probabilities of the first and second kinds are introduced by which the usual probability of error can be computed iteratively.

In the second part, only "noisy" observations are available. In this case, a new model is defined in which the states of the continuous-time Markov chains are described by random processes, but the transition times can be observed. Iterative, optimal (minimum Bayes risk) decision rules are derived and conditions are established under which these rules perform effectively. Optimum-adaptive decision rules are defined when the underlying model is not completely specified. Decision rules are formulated with two types of random processes.

Finally, the situation when the transition times cannot be observed is investigated for the special case in which there are two pattern classes and the states are observed with additive, Gaussian, white noise. Both discrete and continuous observation times are considered. Computationally feasible algorithms are derived for the likelihood ratio which optimally solve the problem, assuming a discrete-sampling scheme. Also, stochastic differential equations are found for the continuous logarithm of the likelihood ratio and the continuous conditional probabilities of errors from discrete results by a limiting operation. The results are applied to the specific problem

of detections

Markov chain

of detecting a random telegraph signal (two-state, continuous-time Markov chain) in white noise.

BAYESIAN DECISION MAKING AND LEARNING FOR CONTINUOUS-TIME MARKOV SYSTEMS

By

Erdal Panayırcı

A THESIS

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

DOCTOR OF PHILOSOPHY

Department of Electrical Engineering and Systems Science

1970

TO

MY MOTHER AND FATHER

The a

encourageme

Science, his

of this rese

ing the pr

well as f

Spe

ment of Sta

concerning t

For

Arthur W.,

Robert Ir.

B. Stapleton,

Demain, Depa

Als

and Systems

his recommen

Final

of Istanbul a

University, f

ACKNOWLEDGEMENTS

The author very gratefully acknowledges the guidance and encouragement of Dr. Richard C. Dubes, Department of Computer Science, his teacher and academic advisor, throughout the course of this research. The author is indebted to Dr. Dubes for suggesting the problem as a potentially fruitful area for research, as well as for his generous devotion of time to provide counsel.

Special thanks are due to Dr. Dennis C. Gilliland, Department of Statistics and Probability, for many useful discussions concerning the problems treated in this thesis.

For their interest and consideration of this research, the author wishes to acknowledge the other members of his doctoral committee: Dr. Carl V. Page, Department of Computer Science, Dr. James H. Stapleton, Department of Statistics and Probability, Dr. Rita Zemach, Department of Electrical Engineering and Systems Science.

Also, Dr. Yılmaz Tokad, Department of Electrical Engineering and Systems Science, deserves a special acknowledgement because of his recommendation and continuous encouragement.

Finally, acknowledgement is made to the Technical University of Istanbul and the Division of Engineering Research, Michigan State University, for the financial support which made this work possible.

TABLE OF CONTENTS

Page

LIST OF FIGURES

LIST OF APPENDICES

LIST OF APPENDICES

N

M

p_s^i

$\xi_s^i(\cdot)$

x^k

t^k

$\xi(x^k, t^k) = s$

LIST OF SYMBOLS

Symbol	Description
N	Number of states in each Markov chain.
M	Number of pattern classes (Markov chains).
P_s^o	Prior probability for pattern class s.
$P_s^o(\cdot)$	Density (probability mass) function for first observation when chain s is active.
x^k	$\triangleq (x_1, x_2, \ldots, x_k)$ Random variables representing state numbers (also, the values for a particular realization of the random variables).
Λ	$= \{1,2,\ldots,N\}$ Range space for each x_i.
t^k	$\triangleq (t_1, t_2, \ldots, t_k)$ Random variables representing sojourn times (also, the values for a particular realization of the random variables).
Q_s	$= [q_{ij}^{(s)}]$ Transition-rate matrix for pattern class s.
$q_i^{(s)}$	$\triangleq -q_{ii}^{(s)}$ Parameter establishing distribution of sojourn in state i when class s is active.
θ	Point in state space (Pattern class).
$g(x^k, t^k \mid \theta = s)$	$\triangleq g_s(x^k, t^k)$ Product of density function for t^k and probability mass function for x^k when class s is active.
$r_{ij}^{(s)}$	$\triangleq q_{ij}^{(s)}/q_i^{(s)}$ Transition probability from state i to state j for the jump chain associated with class s.

Symbol

N_{ij}

Z_i

$\bar{X}_i$

$y_s^{n_s}$

$x_s^{n_s}$

[illegible]

•

c_s

r_s

$f_s(c_s, r_s), \cdot$

$z_{ij}^{(s)}$

$z_i^{(s)}$

Symbol	Description
N_{ij}	$= N_{ij}(x^k)$ Number of one-step transitions from state i to state j in $\mathbf{x}^k$.
Z_i	$= Z_i(t^k)$ Total sojourn time in t^k spent in state i.
K_i	$= K_i(x^k)$ Number of occupancies of state i in x^k.
$y_s^{n_s}$	$\triangleq (y_{s1}, y_{s2}, \ldots, y_{sn_s})$ States from the training sample function produced by pattern class s.
$\tau_s^{n_s}$	$\triangleq (\tau_{s1}, \tau_{s2}, \ldots, \tau_{sn_s})$ Sojourns from training sample function produced by pattern class s.
y	Concatenation of states from all training functions.
τ	Concatenation of sojourns from all training functions.
q_s	$= (q_1^{(s)}, q_2^{(s)}, \ldots, q_N^{(s)})$ Vector of diagonal entries of Q_s.
r_s	Vector of all non-zero terms from the transition matrix of the jump chain for class s.
$f_s(q_s, r_s \mid y, \tau)$	Posterior density for the parameters of the class s, given all the training patterns.
$n_{ij}^{(s)}$	$= n_{is}^{(s)}(y_s^{n_s})$ Number of one-step transitions from state i to state j in the training patterns $y_s^{n_s}$ from class s.
$z_i^{(s)}$	$= z_i^{(s)}(\tau_s^{n_s})$ Total sojourn time in state i recorded in the training pattern $\tau_s^{n_s}$ from class s.

Symbol	Description
$k_i^{(s)}$	$= k_i^{(s)}(y_s^{n_s})$ Number of occupancies of state i in the training pattern $y_s^{n_s}$ from class s.
$f(q_s,r_s\|\tau_o,y_o)$	Prior density for the parameters from class s.
$[w_i^{(s)},v_i]_{i=1}^{N}$, $[\mu_{ij}^{o(s)}]_{i,j=1}^{N}$	Parameters of the prior density for (q_s,r_s).
$[W_i^{(s)},V_i]_{i=1}^{N}$, $[m_{ij}^{(s)}]_{i,j=1}^{N}$	Parameters of posterior density for (q_s,r_s).
w.p.1	With probability one.
Δ_{ij}	Kronecker delta.
P_e	Total probability of error
ρ	Bhattacharyya distance.
α_o,β_o	Probability of error of the first kind and the second kind.
$\alpha_o(k),\beta_o(k)$	Conditional probability of errors of first and second kinds.
$\rho_1,\rho_2,\ldots,\rho_k$	Times at which transitions occur.
$R(t)$	Auto-correlation function.
$\delta(t)$	Impulse function.
Λ_k	Likelihood function.
η	Threshold.
$u(\cdot)$	Unit step function.

In r

have receive

branches of

speech and

machines, cl

tion and fi

research.

In

"computer

arising in

be establis

other sim

consists

which mea

ments, cal

A "patter

is very l

vestigat

linear c

in a give

or order

CHAPTER I

INTRODUCTION

In recent years, pattern recognition and learning theory have received a great deal of attention from investigators in many branches of sciences. Some applications are: Recognition of speech and handwritten characters, checkers and chess playing machines, classification of electrocardiograms and EEG's, detection and filtering theory, system identification and operations research.

In this thesis, "pattern recognition" is another name for "computerized decision making". Given a set of objects, each arising in one of a finite number of sources, an algorithm is to be established which efficiently classifies the objects and all other similar objects. A pattern recognition problem, in general, consists of two sub-problems. The first sub-problem determines which measurements should be taken on the objects. These measurements, called "features", characterize all the possible objects. A "pattern" is a vector of feature measurements. At present, there is very little general theory for the selection of features. Investigators have been concerned with the selection of subsets or linear combinations of existing features or with ordering features in a given set of measurements. The criterion for feature selection or ordering is often based on either the importance of features in

characteriz
to the perf
problem in
or making d
pattern re
"classifie
mathematica
features re
probability
the patter
recogniti
satisfying
minimum pr

Th
must be s
classes are
tributions
the source
patterns".
the case w
of the tra
(learning

A
tion refers
made as to
acterizing
partition

characterizing the patterns or on the contribution of the features to the performance of a recognition algorithm. The second subproblem in pattern recognition is the problem of classification, or making decisions about the source of the patterns. Thus, a pattern recognizer consists of a "feature extractor", and a "classifier". Statistical decision theory provides a powerful mathematical tool for solving the classification problem when the features representing the pattern classes can be described by probability distributions. The application of decision theory to the pattern recognition problem is called "parametric pattern recognition". One can proceed to obtain optimal decision rules satisfying a (subjectively chosen) classification criterion; e.g., minimum probability of misclassification (probability of error).

The problem of learning in parametric pattern recognition must be solved when the distributions characterizing the pattern classes are inadequately known. The unknowns of the class distributions are "learned" from sample patterns drawn from each of the sources, or pattern classes. These samples are called "training patterns". Supervised learning (learning with a teacher) refers to the case when the training patterns are classified. When the origins of the training patterns are unknown, learning is non-supervised (learning without a teacher).

A so-called "non-parametric" approach to pattern recognition refers to the design of classifiers in which no assumption is made as to the form of the underlying probability distributions characterizing each class. The goal of the recognition system is to partition the feature space into regions such that each region can be identif

a discrim

A pattern

largest di

are usuall

set of fu

In

problem is

pattern is

pattern cla

time Mark

chains are

the medi

several f

decision r

decision

learning

The gener

area in

Dobes [1-

filtering

is a fun

1.1 LITE

A

tion and

[N-3]. Pre

be identified with a pattern class. This can be achieved by defining a discriminant function for each pattern class on the feature space. A pattern is classified by choosing the class corresponding to the largest discriminant function. Training patterns from each class are usually available and the problem is to establish a reasonable set of functions from them.

In this thesis, a general parametric pattern recognition problem is investigated in which the classification of an unknown pattern is inferred from a finite set of training patterns. Each pattern class is characterized by a different N-state, continuous-time Markov chain. The stationary transition matrices of the chains are the parameters of the pattern classes. Depending upon the medium (noisy or noiseless) in which the observations are made, several feature selection schemes are considered. Optimum Bayes decision rules are formulated as solutions to problems in statistical decision theory. When the parameters are not known, a supervised learning scheme that uses classified training patterns is employed. The general model considered here finds, specifically, an application area in the classification of EEG's which has been investigated by Dubes [D-5], recently. It is also applicable to detection and filtering problems with Gaussian noise when the underlying signal is a function of a finite dimensional Markov process.

1.1 LITERATURE REVIEW

A comprehensive survey of early works on pattern recognition and learning theory has been written by Nagy [N-1]. Nilsson [N-3] presented some of the theory of "learning machines", or machines of

[3-1] ident

central t

the latest

sequential

theory. F

the [3-

author em

matical m

T

attempt

Fisher [3-

[3-] and

deal with

to classif

F

studied th

question o

better p

true valu

Patrick a

schemes.

Th

marized.

determin

Markov cha

and the tra

nachines which can be trained to recognize patterns, and Sebestyen [S-1] identified the task of finding clustering transformations as central to the design of pattern recognizers. Fu [F-4] presented the latest developments in the area of machine learning, emphasizing sequential methods in statistical decision theory and estimation theory. Recent books edited by Watanable [W-3], Kanal [K-3] and Tou [T-2] are collections of papers on pattern recognition. Each author emphasizes the philosophy of the approach rather than mathematical derivations or experimental data.

The problem of classification has, indeed, received much attention by statisticians. Of the many sources, the books of Fisher [F-2], Anderson [A-1], Raiffa and Schlaifer [R-1], Ferguson [F-2] and Blackwell and Girshick [B-3] should be mentioned that deal with the theory of statistical techniques and the application to classification problems.

For learning theory, Spragins [S-5] and Braverman [B-4] studied the convergence question in supervised learning. This question deals with the sufficient conditions under which the parameter posterior density approaches the delta function about the true value of the parameters as the number of samples increases. Patrick and Hancock [P-1] gave a rather general approach to learning schemes.

The literature closely related to the thesis is now summarized. Dubes and Donoghue [D-4] considered the problem of determining which of a finite set of N-state, discrete-parameter Markov chains is active. The states are observed without noise and the transition probabilities for the chains in question are unknown. B

However, t

The res

from refer

continuous-

transition

constitute

are als

a first

states of

most of

optimum

system and

meters can

and Bartle

the underly

states of

of observ

adaptive

P, and

studied th

when the

meter Mar

variables,

set of

decision

unknown. A Bayesian strategy is employed throughout the report. However, they did not investigate probability of error problems. The results in Chapters II and III presented in this thesis differ from reference [D-4] in that the pattern classes are described by continuous-parameter Markov chains. The unknown parameters are the transition rate matrices (Q-matrices) and noisy observations are considered. Exact and asymptotic probability of error expressions are also derived for several cases.

When the state activity of a general system is described by a first order, homogeneous, discrete-parameter Markov chain and the states of the chain can be observed only in the presence of noise, most of the literature deals with the design of optimum and sub-optimum decision rules for making decisions about the states of the system and establishing conditions under which the unknown parameters can be learned. Billingsley [B-2], Martin [M-2], Good [G-1] and Bartlett [B-1] estimated the transition probability matrix of the underlying chain assuming, by some external means, that the states of the system could be observed. Removing the assumptions of observability of the states, Raviv [R-2] constructed a class of adaptive decision rules using an estimate of the transition matrix, P, and only part of the past observations. Recently, Signori [S-2] studied the problem of determining the optimum-adaptive decision rule when the observations were governed by an underlying discrete-parameter Markov chain. The conditional densities of the observed random variables, given the state of the system, were characterized by a set of unknown parameters. He derived an iterative, optimum adaptive decision rule with the capability of using future and past observations as well as

consistent

of sub[illegible]

[illegible]

the pr[illegible]

terms with

of the par[illegible]

the solut[illegible]

[illegible]

[illegible]

quadrati[illegible]

stochastic

taking pr[illegible]

taking a

a state w[illegible]

[illegible]

[illegible]

property.

[illegible]

probabili[illegible]

nature.

[illegible]

[illegible]

[illegible] have

[illegible]

stochastic

[illegible]

as well as present observations. He also constructed a variety of consistent estimators for the unknown parameters which yield a class of suboptimal rules.

Patrick and Hancock [P-1] gave a rather general approach to the problem of learning for classification and recognition of patterns with or without supervision. Their model for the quantization of the parameter space is given as a reference in Chapter IV, to the solution of the problem of finite computer storage in implementing optimal decision rules and learning schemes.

Hilborn and Lainiotis [H-1] investigated the optimal (in the quadratic sense) nonlinear estimation of discrete-time or sampled stochastic processes, where the processes can be characterized as having probability distributions of known functional form but containing a set of unknown parameters. A Bayes optimal estimate for a state was to be determined and expressed in terms of the parameter-conditional-optimum estimates and another statistic which could be computed recursively. The observations obeyed a generalized Markov property. The results of Chapters IV and V in this thesis differ from [H-1] since the optimization criterion used here was minimum probability of error. The decision problem is alos different in nature.

Recent studies of the problem of detecting an arbitrary random signal in the presence of additive Gaussian noise by Kailath [K-2] have resulted in an exact formula for the optimal likelihood ratio, which applies to the detection of any continuous second order stochastic process. The result must be expressed in terms of continuous stochastic integrals, so it cannot be implemented directly.

McLendon [

arbitrary

approximat

for the

to the s

observati

tained by

Th

taught

Lubanev

quantized,

signal,

Markov

the B

and obtain

Kilman and

a two

They first

for th

cases an

non-linear

thesis, m

with a part

3.0 TH

Th

learning t

McLendon [M-1] investigated the general problem of extracting an arbitrary random process from additive white noise. Under certain approximations, computationally feasible algorithms were derived for the logarithm of the likelihood ratio. The assumption essential to the solution of the problem was that the joint densities of the observation processes, when the signal is present, could be approximated by Gaussian densities (Pseudo Bayes approximation).

The question of detection of Markov processes in a noisy background has been studied by several authors. Nifontov and Likharev [N-2] considered the optimal detection of a Binary, quantized, Markov signal in the presence of noise similar to the signal. Sosolin [S-4] investigated the optimal detection of Gauss-Markov noise with discrete-time observations. They both adopted the Bayes likelihood ratio criterion as an optimum decision rule and obtained recurrent relationships for the likelihood ratio. Kulman and Stratonovic [K-6] provided optimal devices for detecting a random telegraph signal in the presence of white Gaussian noise. They first obtained a non-linear stochastic differential equation for the optimum filtering and then tried to solve it for some special cases and compared the results of the probabilities of errors with non-linear and linear filtering. The work in Chapter V of this thesis, related to their results, was done independently and deals with a particular model which yields more specific results.

1.2 THESIS OUTLINE AND CONTRIBUTIONS

The emphasis of this research is on pattern recognition and learning theory. The first major contribution of the thesis appears

In Chap

adaptive

has not

that

[illegible]

[illegible]

[illegible]

[illegible]

Sect

[illegible]

stat

Chap

[illegible]

is als

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

express

[illegible]

[illegible]

[illegible]

[illegible]

in Chapter II which contains a formal development of optimal and adaptive decision making and learning, employing a new model which has not been heretofore studied. The contribution lies in the fact that, under some necessary and sufficient conditions, a finite set of constant parameters can be found which uniquely defines the underlying model.

Chapter II is devoted to the case of perfect observation (noiseless case). The basic model and the decision problem are defined in Sec. 2.1, Appendix A and Appendix B. In the early sections of the chapter, the optimum and the optimum-adaptive decision rules are found and expressed in terms of sufficient statistics of finite dimensions for two cases. A supervised learning scheme is employed to learn the paramters, and the existence of reproducing prior densities for them is exhibited. It is also shown that the computer memory needed to implement these rules is fixed. Finally, a computer simulation is developed and discussed for a special case.

The second major contribution appears in Chapter III in which the probability of error is studied. Some specific results are obtained for the model considered in Chapter II. In the case considered, all quantities in the model are known and there are only two pattern classes. In Sec. 3.3, exact probability of error expressions are derived for a particular case while in the following sections, upper and lower bounds and asymptotic expressions are obtained. When the number of observations increases without bound, the bound on the probability of error is shown to approach zero. In the last section of the chapter, conditional error probabilities

of the fir
the total
derived f

3.
optimal an
are invest
this prob
which the
The main a
The state
sampling s
is derive
assuming a
results ar
of Chapter
When the
otherwise
The strate
[illegible]

4.
In the a
moves. A
assumed.
[illegible]
two pat
[illegible]
[illegible]

of the first kind and the second kind are introduced from which the total probability of error can be computed. The expressions derived for the error probabilities are original and self-contained.

Final contributions appear in Chapters IV and V in which optimal and adaptive decision-making and parameter-learning problems are investigated under an imperfect observation mechanism. Inserting this condition into the model of Chapter II requires a new model in which the states of the chains are described by random processes. The main assumption of Chapter IV is that the transition times from one state to another can be observed. After discussing several sampling schemes for selecting features, the optimum decision rule is derived and its basic components are generated iteratively, assuming all the parameters of the model are known. Analytic results are obtained for two special cases. In the second part of Chapter IV, adaptive decision-making and learning are studied when the model is not completely specified. A theorem about the convergence of the optimum-adaptive decision rule is provided. The storage problem in implementing the adaptive decision rule and supervised learning algorithm is discussed.

In Chapter V, the model and decision problem are studied but the assumption of observability of the transition times is removed. A uniform sampling scheme (discrete-time observations) is assumed. The Bayes likelihood algorithm is developed for the optimum decision rule and the recurrent expressions for the likelihood ratio is derived. In the case of continuous observations, non-linear stochastic differential equations are derived for the logarithm of the likelihood ratio and the conditional error probabilit

IV and V ar

present in

probabilities of the first and second kinds. The results of Chapters IV and V are original and have not appeared in the literature and present one of the major contributions of the thesis.

CHAPTER II

DECISION MAKING AND LEARNING WITH OBSERVABLE STATES AND TRANSITION TIMES

The basic problem considered here is that of determining which of a finite set of M continuous-time, discrete-state Markov process is active, based upon observations of sample functions, when the stationary transition probability matrices, $[P_{ij}^{(s)}(t)]_{i,j=1}^{N}$; $s \in \{1,2,\ldots,M\}$, for the processes in question are unknown. The main results of this chapter rely heavily upon the definitions and theorems related with continuous-time, discrete-state Markov processes (continuous-parameter Markov chains) which are given without proofs in Appendix A.

A Bayesian strategy is employed throughout. The problem is formulated as a problem in statistical decision theory. Prior distributions which lead to convenient computer implementations are chosen for the unknown parameters. The amount of computer storage required is of prime importance.

The decision problem is defined and a source model is chosen for generating observations in Sec. 2.1. In Sec. 2.2, observation and parameter spaces are defined, and in Sec. 2.3, the optimal decision rule is derived when the transition rate matrices (Q-matrices) are known for every pattern class. In Sec. 2.4, adaptive decision rules are derived when the transition-rate matrices are not known. A supervised learning scheme is employed to learn the unknown to

prior den

ing trans

form for

in Sec. 2.

of a spe

summation

2.2 S

B

which

decisi

Markov

characteri

"feature

numbers

the pr

observat

the tr

used

are emi

vation pr

continuous

conditi

determine

A

During the

unknown transition-rate matrices and the existence of reproducing prior densities is demonstrated in Sec. 2.5. Algorithms for learning transition rates are introduced in Sec. 2.6, while the final form for the optimal optimum-adaptive Bayes decision rule is obtained in Sec. 2.7. Section 2.8 is assigned to the computer implementation of a specific problem. Finally, the main results of the chapter are summarized in Sec. 2.9.

2.1 SOURCE MODEL

Before going into decision rules and learning, the model by which observations are generated must be chosen. The object of all decision rules is to decide which of M continuous-parameter Markov chain is active; each continuous-parameter Markov chain characterizes a pattern class. The observable quantities, or the "features", are the sojourn times in the states and the state numbers themselves. The transition-rate, or Q, matrices defining the processes are the parameters of the distributions governing the observations and must be learned from the training data. The training data consist of sample functions from labelled sources and are used to form posterior densities for the parameters which, in turn, are employed in Bayes decision rules. The properties of the observation processes, some important definitions and theorems about continuous-parameter Markov chains, and the necessary and sufficient conditions under which infinitesimal parameters can uniquely determine a process are given in Appendix A.

A key assumption is that a single Markov chain is active during the entire observation interval. A decision about the identity

function.

manner

in Appendix

Suppose a

a waiting

$q_i > 0$,

and if

only

rand

parameter

probability

distrib

process

function

Fig. A.1.

[illegible]

in the

of values

time rand

values on

on the spa

is the

the rand

identity of that chain is to be made after observing a sample function. Each sample function is assumed to be generated in a manner consistent with the assumptions and conditions explained in Appendix A. The generation process can be pictured as follows: Suppose a typical realization starts in any state, say i. Then, a waiting time, having an exponential distribution with parameter $q_i > 0$, determines the length of time spent in state i. At the end of this sojourn, the process jumps to state j with probability $\frac{q_{ij}}{q_i}$, $j \neq i$. The process stays in the new state for a random duration, as determined by an exponential distribution with parameter q_j, and then moves to another state, say k, with probability $\frac{q_{jk}}{q_j}$, $k \neq j$. The sojourn time again has an exponential distribution with parameter q_k. All possible realization of the process can be generated by this procedure. A typical sample function constructed by the above procedure is illustrated in Fig. A.1.

2.2 DECISION RULES

The observation space is defined first. Each random variable in the sequence $x^k \triangleq (x_1,x_2,\ldots,x_k)$ of state random variables takes on values in a finite space $\Lambda = \{1,2,\ldots,N\}$, $N < \infty$, and each sojourn time random variable in the sequence $t^k \triangleq \{t_1,t_2,\ldots,t_k\}$ takes on values on the positive real line. All random variables are defined on the space $S_k = \{\omega\colon \omega \in \Lambda^k \times [0,\infty)^k\}$. Thus, the sample space S_k is the 2k-dimensional Euclidean space of all sequences $\omega = (\prod_{i=1}^{k} \xi_i) \times (\prod_{i=1}^{k} \eta_i)$, where $\xi_i \in \Lambda$, $\eta_i \in [0,\infty)$, $\forall i$. In particular, the random variable x_i, $i \leq k$, is defined as $X_i(\omega) = \xi_i$ and the random var

Th

given in A

discrete-

defined as

Using [3-1,

with the

The

Percent ran

measure out

=

where A

each set

distribut

this model

conditional

Thus, ^{K} w

The

discrete d

measure

the status

Then, the

random variable t_i, $i \le k$, is defined as $t_i(\omega) = \eta_i$.

The necessary and sufficient conditions and Theorem A.1 given in Appendix A show that $\{x_k;\ k \ge 1\}$ is a discrete-state, discrete-time Markov chain with transition matrix $[r_{ij}]_{i,j=1}^{N}$ defined as follows:

$$r_{ij} = \begin{cases} q_{ij}/q_i & j \ne i \\ 0 & j = i \end{cases} \tag{2.2.1}$$

Chung [C-2] calls such a Markov chain, the jump chain associated with the continuous-parameter Markov chain, $\{x_t;\ 0 \le t < \infty\}$.

The sojourn times $(t_1,t_2,\ldots,)$ are conditionally independent random variables. That is, if $P(\cdot)$ is a probability measure defined on the sample space S_k, then

$$\begin{aligned} &P(t_1 \in A_1, t_2 \in A_2, \ldots, t_k \in A_k | x_1 = i, x_2 = j, \ldots, x_k = \ell) \\ &= P(t_1 \in A_1 | x_1 = i) \ldots P(t_k \in A_k | x_k = \ell) \end{aligned} \tag{2.2.2}$$

where A_i is suitable Borel set on R^1, $i = 1,2,\ldots,k$. Hence, for each sequence of realizations $x^k = (x_1,x_2,\ldots,x_k)$, a conditional distribution for $t^k = (t_1,t_2,\ldots,t_k)$, given x^k is defined. In this model, the conditional distribution can be described by the conditional density $f(t^k|x^k) \triangleq f(t_1|x_1)f(t_2|x_2)\ldots f(t_k|x_k)$ on $[0,\infty)^k$ with respect to Lebesgue measure μ^k.

The mass function $p(x^k) \triangleq p(x_1,x_2,\ldots,x_k)$, which is a discrete density over N^k points in R^k with respect to counting measure ν^k is defined in terms of the initial distribution over the states and the transition probability matrix, $[r_{ij}]_{i,j=1}^{N}$. Then, the joint density for (x^k,t^k), denoted by $g(\cdot)$, is well-defined in

duct of c

expressed

a

Th

derivati

arbitrary

following

derived f

a special

2.3 [illegible]

Th

special

The optimal

and is giv

It follows

[illegible]

defined in terms of $f(t^k|x^k)$ and $p(x^k)$, with respect to the product of counting and Lebesgue measures $\nu^k x \mu^k$ on R^{2k} and can be expressed as

$$g(x^k,t^k) \overset{\Delta}{=} g(x_1,\ldots,x_k, t_1,\ldots,t_k)$$
$$= f(t_1,\ldots,t_k|x_1,\ldots,x_k)p(x_1,\ldots,x_k) \qquad (2.2.3)$$

The general elements of the decision making problem and derivation of the non-randomized, optimal decision rules for any arbitrary loss function, $L(.,.)$ are given in Appendix B. In the following two sections, optimal and adaptive decision rules are derived for the cases of known and unknown Q-matrices, assuming a special loss function.

2.3 OPTIMAL DECISION RULES WHEN Q-MATRICES ARE KNOWN

The first case to be considered assumes that the Q-matrix, $[q_{ij}^{(s)}]_{i,j=1}^{N}$ is known for every pattern class, $s = 1,2,\ldots,M$. The special "0-1" loss function is chosen, defined by $L(i,j) = 1 - \Delta(i,j)$[1]. The optimal decision rule, $d^*(.)$, follows from (B.5), Appendix B, and is given by the Bayes decision rule:

$$d^*(x^k,t^k) = s \quad \text{if} \quad s \quad \text{is the first index such that}$$
$$P(\theta = s|x^k,t^k) \geq P(\theta = \ell|x^k,t^k),$$
$$\forall\ \ell \neq s,\ \ell,s \in \Lambda \qquad (2.3.1)$$

It follows from Bayes rule that

[1] $\Delta(i,j)$ denotes the Kronecker delta.

F

Defining

(2.3.2),

c^*

F_s

T

$F_s(x^*, t$

T

optimal

F

$=$

The second

$F_s(x_1, x$

Then, the

(2.3.6) as

$F_s(x^k, t^k) =$

$$P(\theta = s|x^k,t^k) = \frac{P(\theta=s)g(x^k,t^k|\theta=s)}{g(x^k,t^k)} \,. \tag{2.3.2}$$

Defining $g_s(x^k,t^k) \triangleq g(x^k,t^k|\theta = s)$, $P_s^o = P(\theta = s)$ and employing (2.3.2), (2.3.1) becomes

$$\begin{aligned} &d^*(x^k,t^k) = s \quad \text{if} \quad s \quad \text{is the first index such that} \\ &P_s^o \; g_s(x^k,t^k) \geq P_\ell^o \; g_\ell(x^k,t^k) \;\; \forall \; \ell \neq s, \; s,\ell \in \Lambda. \end{aligned} \tag{2.3.3}$$

The density functions required in (2.3.3) are given by

$$g_s(x^k,t^k) = f_s(t_1,t_2,\ldots,t_k|x_1,x_2,\ldots,x_k)\,p_s(x_1,\ldots,x_k) \qquad \forall \; s \in \Lambda \tag{2.3.4}$$

The first factor in (2.3.4) is now computed. From the conditional independence of $(t_1,t_2,\ldots,t_k)$ and (A.14),

$$\begin{aligned} f_s(t_1,\ldots,t_k|x_1,\ldots,x_k) &= \prod_{i=1}^{k} f_s(t_i|x_i) \\ &= \left(\prod_{i=1}^{k} q_{x_i}^{(s)}\right) \exp\left\{-\sum_{i=1}^{k} q_{x_i}^{(s)} t_i\right\} \end{aligned} \tag{2.3.5}$$

The second factor in (2.3.4) is found from (A.13).

$$p_s(x_1,x_2,\ldots,x_k) = P_s^o(x_1) \prod_{i=1}^{k-1} \frac{q_{x_i,\,x_{i+1}}^{(s)}}{q_{x_i}^{(s)}} \tag{2.3.6}$$

Then, the joint density $g_s(x^k,t^k)$ is given from (2.3.5) and (2.3.6) as:

$$g_s(x^k,t^k) = \left(\prod_{i=1}^{k} q_{x_i}^{(s)}\right) \exp\left\{-\sum_{i=1}^{k} q_{x_i}^{(s)} t_i\right\} P^o(x_1) \prod_{i=1}^{k-1} \frac{q_{x_i,\,x_{i+1}}^{(s)}}{q_{x_i}^{(s)}} \tag{2.3.7}$$

E

as follow

transitio

$K_i = K_i(x^i$

not count

times in

sojourn t

Then, (2.3

$g_s(x^k, t^k$

Th

the optima

form as g

2.4 ADAPTI

Wh

are not kn

Equation (2.3.7) can be written in a more convenient way as follows: Let $N_{ij} = N_{ij}(x^k)$ denote the number of one-step transitions from state i to state j in $x^k = (x_1,\ldots,x_k)$. Let $K_i = K_i(x^k)$ be the number of occupancies of state i in x^k, not counting x_1, and let $t_{ij} = t_{ij}(t^k)$ be the waiting (sojourn) times in state i, for the jth occupancy of the process. The total sojourn time in state i, $Z_i = Z_i(t^k)$ is then,

$$Z_i = t_{i1} + t_{i2} + \ldots + t_{iK_i} \qquad \forall\, i \in \Lambda$$

Then, (2.3.7) can be written as,

$$g_s(x^k,t^k) = \left[\prod_{i=1}^{N} (q_i^{(s)})^{K_i}\right] \exp\left\{-\sum_{i=1}^{N} q_i^{(s)} Z_i\right\} P_s^o(x_1) \prod_{i=1}^{N} \prod_{\substack{j=1\\ j\neq i}}^{N} \left(\frac{q_{ij}^{(s)}}{q_i^{(s)}}\right)^{N_{ij}} \tag{2.3.8}$$

Taking natural logorithms of (2.3.8) and noting that

$$\begin{aligned} \sum_{\substack{j=1\\ j\neq i}}^{N} N_{ij} &= K_i \quad \text{if} \quad x_k \neq i \\ &= K_i + 1 \quad \text{if} \quad x_k = i \end{aligned} \tag{2.3.9}$$

the optimal decision rule in (2.3.3) can be expressed in a simpler form as follows:

$$d^*(x^k,t^k) = s \quad \text{if} \quad s = \ell \quad \text{maximizes}$$

$$\left\{\ell n\left[P_\ell^o(x_1) P_\ell^o\, q_{x_1}^{(\ell)}\right] - \sum_{i=1}^{N} q_i^{(\ell)} Z_i + \sum_{i=1}^{N} \sum_{\substack{j=1\\ j\neq i}}^{N} N_{ij}\, \ell n q_{ij}^{(\ell)}\right\} \tag{2.3.10}$$

2.4 ADAPTIVE DECISION RULES WHEN Q-MATRICES ARE NOT KNOWN

When the infinitesimal parameters, $\{q_{ij}^{(s)}\}_{i,j}^{N}$, $s \in \{1,2,\ldots,M\}$, are not known, they are treated as parameters in density functions describing

such cond

extracted

sojourn t.

observed s

the decis.

ing data :

$(y_s^{n_s}, \ell_s^{n_s}$

and if s

$d^*(y^*$

$p_s^n\,g$

where

$g_s(y^k, \ell^k$

where

The term

The

pattern cla

from the g

of (2.4.2)

describing the observations. To make effective decisions under such circumstances, information about these unknowns must be extracted from classified observations. The state numbers and the sojourn times of sample functions from each pattern class are observed and the infinitesimal parameters can be eliminated from the decision rule using supervised learning techniques. The training data from pattern class s form a random vector $(y_s^{n_s}, \tau_s^{n_s}) \triangleq (y_{s1}, y_{s2}, \ldots, y_{sn_s}, \tau_{s1}, \tau_{s2}, \ldots, \tau_{sn_s})$ of states numbers and of sojourn times. The optimal decision rule in this case is

$$d^*(x^k, t^k) = s \quad \text{if} \quad s \quad \text{is the first index such that}$$

$$P_s^o \, g_s(x^k, t^k | y_s^{n_s}, \tau_s^{n_s}) \geq P_\ell^o \, g_\ell(x^k, t^k | y_\ell^{n_\ell}, \tau_\ell^{n_\ell})$$

$$\forall \, \ell \neq s, \ \ell, s \in \Lambda \tag{2.4.1}$$

where

$$g_s(x^k, t^k | y_s^{n_s}, \tau_s^{n_s}) = \int_{R^{N^2}} g_s(x^k, t^k | y_s^{n_s}, \tau_s^{n_s}, q_s, r_s) f(q_s, r_s | y_s^{n_s}, \tau_s^{n_s}) dq_s dr_s \tag{2.4.2}$$

where

$$q_s \triangleq (q_1^{(s)}, q_2^{(s)}, \ldots, q_N^{(s)});$$

$$r_s = \{r_{ij}^{(s)}\}_{i,j=1}^N \tag{2.4.3}$$

The term $r_{ij}^{(s)}$ was defined in (2.2.1); $\sum_{\substack{j=1 \\ j \neq i}}^N r_{ij}^{(s)} = 1, \forall i \in \Lambda$.

The supervised learning procedure is the same for all pattern classes so the subscript and superscript s will be dropped from the following development. The first factor in the integrand of (2.4.2) can be written in a form similar to (2.3.8) as follows:

$g(x^k, t^k$

where N

as before

in Sec. 2.

$g(x^k, t^k$

E

function

tion. As

whenever a

(ξ_s, τ_s).

2.5 SUM

T

function

learning

shown that

subscript

Th

a reprodu

statistic

Spragins

statistic,

$$g(x^k,t^k|y^n,\tau^n,q,r) = \left(\prod_{i=1}^{N} q_i^{K_i}\right) \exp\left\{-\sum_{i=1}^{N} q_i Z_i\right\} P^o(x_1) \prod_{i=1}^{N} \prod_{\substack{j=1\\ j\neq i}}^{N} (r_{ij})^{N_{ij}} \tag{2.4.4}$$

where $N_{ij} = N_{ij}(x^k)$, $Z_i = Z_i(t^k)$ and $K_i = K_i(t^k)$ are defined as before. The second factor in the integrand of (2.4.2) is computed in Sec. 2.5. As a result, the required density has the form:

$$g(x^k,t^k|y^n,t^n) = P^o(x_1) \int_q\int_r \left[\left(\prod_{i=1}^{N} q_i^{K_i}\right) \exp\left\{-\sum_{i=1}^{N} q_i Z_i\right\} \times \prod_{i=1}^{N} \prod_{\substack{j=1\\ j\neq i}}^{N} (r_{ij})^{N_{ij}}\right] f(q,r|y^n,\tau^n)dq\ dr \tag{2.4.5}$$

Equation (2.4.5) shows explicitly that the required density function is proportional to a joint moment of the posterior distribution. As shown in the next section, the computer storage is limited whenever a natural conjugate family of distribution is used for each (q_s,r_s).

2.5 SUPERVISED LEARNING

The object of this section is to form the posterior density function for (q_s,r_s) from the training samples $(y_s^{n_s},\tau_s^{n_s})$. The learning is supervised because the continuous-parameter Markov chain that produces each set of training data is labelled. The s subscript and superscript will be dropped for convenience.

The necessary and sufficient condition for the existence of a reproducing prior distribution for (q,r) is that a sufficient statistic of finite dimension exists for (q,r) (Theorem 2, Spragins [S-6]). To demonstrate this finite-dimensioned sufficient statistic, the likelihood function $g(y^n,\tau^n|q,r)$ can be written from (2.3

$g(y^n,\phi^n$

where n

Ei

exponentia

statisti

$T(y^n,\phi^n$

turn the

T

T

where $(\xi,$

following

$f(\xi,\tau|y_j,\phi$

where $(y$

and $\phi(\xi,$

that the

$I(\xi,\tau) \equiv$

(ξ,τ). T

matrix be

$f(\xi,\tau|y_j,\phi$

from (2.3.8) as:

$$g(y^n,\tau^n|q,r) = \left(\prod_{i=1}^{N} q_i^{k_i}\right) \exp\left\{-\sum_{i=1}^{N} q_i z_i\right\} P^o(y_1) \prod_{i=1}^{N} \prod_{\substack{j=1\\ j\neq i}}^{N} (r_{ij})^{n_{ij}} \qquad (2.5.1)$$

where $n_{ij} = n_{ij}(y^n)$; $z_i = z_i(\tau^n)$, $k_i = k_i(y^n)$.

Equation (2.5.1) shows that $g(y^n,\tau^n|q,r)$ belongs to an exponential family of distributions. Thus, (q,r) has a sufficient statistic of finite dimension. One such statistic is denoted by $T(y^n,\tau^n)$ and can be easily determined by applying the factorization theorem to (2.5.1).

$$T(y^n,\tau^n) = \left(\left\{n_{ij}(y^n)\right\}_{\substack{i,j=1\\ j\neq i}}^{N}, \left\{z_i(\tau^n)\right\}_{i=1}^{N}, \left\{k_i(y^n)\right\}_{i=1}^{N}\right) \qquad (2.5.2)$$

Thus, a reproducing prior density exists for the parameter (q,r). Any reproducing prior density can be written in the following form:

$$f(q,r|y_o,\tau_o) = \frac{g(y_{-m},\ldots,y_o,\tau_{-m},\ldots,\tau_o|q,r)\psi(q,r)}{\int_{q\times r} g(y_{-m},\ldots,y_o,\tau_{-m},\ldots,\tau_o|q,r)\psi(q,r)dq\, dr} \qquad (2.5.3)$$

where $(y_o,\tau_o) \overset{\Delta}{=} (y_{-m},\ldots,y_o,\tau_{-m},\ldots,\tau_o)$ are fictitious observations and $\psi(q,r)$ is an arbitrary positive function of (q,r) except that the denominator in (2.5.3) must exist. In particular, setting $\psi(q,r) \equiv 1$, the following reproducing prior density is obtained for (q,r). This density is the product of N gamma censities and a matrix beta density.

$$f(q,r|y_o,\tau_o) = \left[\prod_{i=1}^{N} \frac{w_i^{v_i+1} q_i^{v_i} e^{-w_i q_i}}{\Gamma(v_i+1)}\right] \cdot \left[\prod_{i=1}^{N} \prod_{\substack{j=1\\ j\neq i}}^{N} B_N(\vec{\mu}_i^o) r_{ij}^{\mu_{ij}^o - 1}\right]. \qquad (2.5.4)$$

where

and $\underline{y}$ =

in all th

to be:

$f(t,r$

where

V_i =

The poster

density f

(2.5.6) is

lated by

conditions

where

$$B_N(\vec{\mu}^{\,o}_i) \triangleq \frac{\Gamma(\Omega_i)}{\prod\limits_{\substack{j=1 \\ j\neq i}}^{N} \Gamma(\mu^o_{ij})} \; ; \quad \Omega_i \triangleq \sum_{j=1}^{N} \mu^o_{ij} \tag{2.5.5}$$

The parameters of this distribution are the matrix $\mu_o = [\mu^o_{ij}]^N_{\substack{i,j=1 \\ j\neq i}}$ where $\mu^0_{ii} = 0$, and the vectors $\underline{v} = (v_1,v_2,\dots,v_N)$ and $\underline{w} = (w_1,w_2,\dots,w_N)$. The posterior density for (q,r), based on all training samples $(y^n,\tau^n) = (y_1,\dots,y_n,\tau_1,\dots,\tau_n)$ is found to be:

$$f(q,r|y^n,\tau^n) = \frac{g(y^n,\tau^n|q,r)f(q,r|y_o,\tau_o)}{\int\limits_{qxr} g(y^n,\tau^n|q,r)f(q,r|y_o,\tau_o)dq\ dr}$$

$$= \left[\prod_{i=1}^{N} \frac{W_i^{V_i+1} q_i^{V_i} e^{-W_i q_i}}{\Gamma(V_i + 1)}\right] \cdot \left[\prod_{i=1}^{N} \prod_{\substack{j=1 \\ j\neq i}} B_N(\vec{m}_i) r_{ij}^{m_{ij}-1}\right]. \tag{2.5.6}$$

where

$$V_i \triangleq v_i + k_i, \quad W_i \triangleq w_i + z_i, \quad m_{ij} \triangleq \mu^o_{ij} + n_{ij}$$

$$m \triangleq [m_{ij}]^N_{i,j=1}; \; B(\vec{m}_i) \triangleq \frac{\Gamma(M_i)}{\prod\limits_{i=1}^{N} \Gamma(m_{ij})}, \; M_i \triangleq \sum_{j=1}^{N} m_{ij} \tag{2.5.7}$$

The posterior density has the same mathematical form as the prior density for (y^n,τ^n). The only difference between (2.5.4) and (2.5.6) is in the parameters of the two densities, which are related by (2.5.7).

The convergence question in supervised learning deals with conditions under which the joint prior densities of the parameters,

$c_s = [c_i^s$

centered

training

this count

THE REM I

A.

F

I.

I.

Then, c_s a

function

F

the first

exhibit th

the unkn

[C-3] is

exists, t

of this s

maximum l

where n_i

for the pa

there exis

the param

statistics

$q_s = \{q_i^{(s)}\}_{i=1}^N$ and $r_s = \{r_{ij}^{(s)}\}_{\substack{i,j=1 \\ j \neq i}}^N$ approach delta functions centered at the true values of the parameters as the number of training patterns increases. The most general theorem for answering this question is given by Spragin's [S-6], and is stated below.

THEOREM 2.5.1 Assume that the following conditions are satisfied.

A. The data generation model under supervised learning is employed; θ_o is the "true" value of θ.

B. $f_\theta(z) > 0$ for all z in a sphere containing θ_o.

C. The posterior density $f_\theta(\cdot|y^n)$ is defined as before.

D. A consistent sequence of estimators $t_1 = t_1(y_1)$, $t_2 = t_2(y_1,y_2),\ldots,t_k = t_k(y_1,\ldots,y_k)$ exists that converges to θ_o wpl.

Then, $f_\theta(z|y^n) \xrightarrow{n\to\infty} \delta(z - \theta_o)$ wpl, where $\delta(\cdot)$ is an impulse function having the same dimension as θ_o.

For the problem considered above, it is easily seen that the first three conditions are satisfied for $g(q,r|y^n,\tau^n)$. To exhibit the existence of a sequence of consistent estimators for the unknown parameters, (q,r), the following well known result [C-3] is used: If a sufficient statistic of an unknown parameter exists, then any maximum likelihood estimator will be a function of this sufficient statistic. Bartlett [B-1] first showed that the maximum likelihood estimators, $\hat{r}_{ij}$, obtained by $\hat{r}_{ij} = n_{ij}/n_i$, where $n_i = \sum_{j=1}^N n_{ij}$ form the consistent sequence of estimation for the parameters r_{ij} in (2.5.1). It can be also shown that there exists a set of consistent maximum likelihood estimators for the parameters q_i in (2.5.1), given in terms of the sufficient statistics k_i and z_i as $\hat{q}_i = (k_i + 1)/z_i$, $i = 1,2,\ldots,N$.

Thus, the

$g(\epsilon, r, y^n,$

2.6 LEA

S

have bee

or to det

training

densities

Here, q^n

posteri

c

define

:

$(y^n, \bullet^n,$

required

Eq

The poste

formula:

Thus, the last condition of Theorem 2.5.1 is satisfied for $g(q,r|y^n,\tau^n)$ and $g(q,r|y^n,\tau^n) \xrightarrow{n\to\infty} \delta(r - r_o, q - q_o)$ w.p.l.

2.6 LEARNING $\{q_{ij}\}_{\substack{i,j=1\\ j\neq i}}^{N}$

So far, only the parameters $\{q_i\}_{i=1}^{N}$ and $\{r_{ij}\}_{\substack{i,j=1\\ j\neq i}}^{N}$ have been learned. It is also possible to learn $\{q_{ij}\}_{\substack{i,j=1\\ j\neq i}}^{N}$, or to determine the posterior densities for the q_{ij}'s, given the training samples, (y^n,τ^n), and to prove that these posterior densities converge to $\delta(q_{ij} - q_{ij}^o)$ w.p.l., as n tends to infinity. Here, q_{ij}^o is the true value of q_{ij}. To find $\psi(q_{ij}|y^n,\tau^n)$, the posterior density for q_{ij}, let

$$q_{ij} = q_i r_{ij} \qquad i,j \in \Lambda (j \neq i) \quad \text{and} \tag{2.6.1}$$

define
$$N_i \triangleq \sum_{\substack{j=1\\ j\neq i}}^{N} n_{ij}(y^k). \tag{2.6.2}$$

The joint posterior density function for (q_i,r_i), given (y^n,τ^n), can be obtained easily by integrating (2.5.6). The required density has the form:

$$f(q_i,r_{ij}|y^n,\tau^n) = \frac{W_i^{V_i+1} q_i^{V_i} e^{-q_i W_i}}{\Gamma(V_i + 1)}$$

$$\times \frac{\Gamma(N_i)}{\Gamma(N_i - n_{ij})\Gamma(n_{ij})} r_{ij}^{n_{ij}-1} (1-r_{ij})^{N_i-n_{ij}-1}$$

$$0 < q_i < \infty; \quad 0 < r_{ij} < 1$$

$$= 0 \quad \text{elsewhere} \tag{2.6.3}$$

The posterior density $\psi(.|y^n,\tau^n)$ is given by the following formula:

Using th

In the

Since

the

in (2.4

Using (2.

where

$$\psi(q_{ij}|y^n,\tau^n) = \int_0^1 \frac{1}{r_{ij}} f(r_{ij}, \frac{q_{ij}}{r_{ij}})\, dr_{ij}$$

$$= \frac{1}{q_{ij}} \int_{q_{ij}}^{\infty} \frac{1}{u} f(\frac{q_{ij}}{u}, u) du$$

$$= \frac{q_{ij}^{n_{ij}-1}}{q_{ij}} \frac{\Gamma(N_i)}{\Gamma(n_{ij})\Gamma(N_i - n_{ij})} \frac{W_i^{V_i+1}}{\Gamma(V_i+1)} \int_{q_{ij}}^{\infty} \frac{u^{V_i}}{u^{n_{ij}}}$$

$$\times \frac{1}{u^{N_i-n_{ij}-1}} (u-q_{ij})^{N_i-n_{ij}-1} e^{-uW_i} du$$

Using the Binomial expansion for $(u - q_{ij})^{\ell_i}$, $\ell_i \triangleq N_i - n_{ij} - 1$, in the above

$$\psi(q_{ij}|y^n,\tau^n) = q_{ij}^{n_{ij}-2} \frac{\Gamma(N_i)}{\Gamma(n_{ij})\Gamma(N_i-n_{ij})} \frac{W_i^{V_i+1}}{\Gamma(V_i+1)} \sum_{k=0}^{\ell_i} (-1)^k \binom{\ell_i}{k} q_{ij}^k \int_{q_{ij}}^{\infty} u^{m_i} e^{-uW_i} du. \qquad (2.6.4)$$

Since $m_i \triangleq V_i - n_{ij} - k$ is a positive integer $k = 0,1,\ldots,\ell_i$, the following integral formula can be used to evaluate the integral in (2.6.4).

$$\int_{q_{ij}}^{\infty} u^{m_i} e^{-uW_i} du = \frac{\Gamma(m_i + 1)}{W_i^{\ell_i + 1}} e^{-q_iW_i} \sum_{r=0}^{m_i} \frac{(q_{ij}W_i)^r}{r!} \qquad (2.6.5)$$

Using (2.6.5) in (2.6.4)

$$\psi(q_{ij}|y^n,\tau^n) = \sum_{k=0}^{\ell_i} \sum_{r=0}^{m_i} a_{kr} q_{ij}^{n_{ij}+r+k-2} e^{-q_{ij}W_i} \qquad q_{ij} > 0$$

$$= 0 \quad \text{elsewhere} \qquad (2.6.7)$$

where

$$a_{kr} \triangleq (-1)^k \begin{pmatrix} N_i-n_{ij}-1 \\ k \end{pmatrix} \frac{\Gamma(N_i)}{\Gamma(n_{ij})\Gamma(N_i-n_{ij})} \frac{\Gamma(V_i-n_{ij}-k+1)}{\Gamma(V_i+1)\Gamma(r+1)} W_i^{n_{ij}+r+k}$$

The fact that $\psi(q_{ij}|y^n,\tau^n) \xrightarrow{n\to\infty} \delta(q_{ij} - q^o_{ij})$ w.p.1., is shown in Appendix C.

2.7 OPTIMUM-ADAPTIVE DECISION RULE

In the previous section, joint posterior density functions for the parameters of each pattern class were obtained and used to eliminate the unknown infinitesimal parameters, $\{q^s_{ij}\}$, from the decision rule. The decision rule obtained in this manner is an adaptive decision rule which is optimal for the prior information and cost function given. Furthermore, the decision rule adapts or converges to what the optimal rule would be if the true parameter values were known. Thus, in summary, the optimum-adaptive Bayes decision rule is given by:

$$d^*(x^k,t^k) = s \quad \text{if} \quad s \quad \text{is the first index such that}$$

$$P^o_s g_s(x^k,t^k|y_s^{n_s},\tau_s^{n_s}) = \max_{\ell\in\{1,2,\dots,M\}} \{P^o_\ell g_\ell(x^k,t^k|y_\ell^{n_\ell},\tau_\ell^{n_\ell})\} \tag{2.7.1}$$

where

$$g_\ell(x^k,t^k|y_\ell^{n_\ell},\tau_\ell^{n_\ell}) = \int_{q\times r} g_\ell(x^k,t^k|y_\ell^{n_\ell},\tau_\ell^{n_\ell},q,r)f(q,r|y_\ell^{n_\ell},\tau_\ell^{n_\ell})dq\ dr \quad \forall \ell \tag{2.7.2}$$

The first and second factors in the integrand above are given in (2.4.4) and (2.5.6), respectively. After some algebra,

$$g_\ell(x^k,t^k|y_\ell^{n_\ell},\tau_\ell^{n_\ell}) = \prod_{i=1}^{N} \frac{\left(W_i^{(\ell)}\right)^{V_i^{(\ell)}+1}}{\left(W_i^{(\ell)}+Z_i\right)^{V_i^{(\ell)}+K_i+1}} \cdot \frac{\Gamma(V_i^{(\ell)}+K_i+1)}{\Gamma(V_i^{(\ell)}+1)}$$

$$\times\ P^o_\ell(x_1) \prod_{i=1}^{N} \frac{\prod_{\substack{j=1\\ j\neq i}}^{N} (N_{ij} + m_{ij}^{(\ell)} - 1)\dots(m_{ij}^{(\ell)})}{(N_i + m_i^{(\ell)} - 1)\dots(m_i^{(\ell)})} \tag{2.7.3}$$

where

$$W_i^{(\ell)} \triangleq w_i^{(\ell)} + z_i(\tau_\ell^{n_\ell});\; V_i^{(\ell)} \triangleq v_i^{(\ell)} + k_i(y_\ell^{n_\ell}),\; m_{ij}^{(\ell)} \triangleq n_{ij}(y_\ell^{n_\ell}) + \mu_{ij}^{o(\ell)}$$

$$Z_i \triangleq Z_i(t^k),\quad K_i \triangleq K_i(x^k),\quad N_{ij} \triangleq N_{ij}(x^k),$$

$$m_i^{(\ell)} \triangleq \sum_{\substack{j=1\\ j\neq i}}^{N} m_{ij}^{(\ell)},\quad N_i \triangleq \sum_{\substack{j=1\\ j\neq i}}^{N} N_{ij} \qquad \forall \ell$$

2.8 COMPUTER IMPLEMENTATION

The optimum-adaptive Bayes decision rule obtained in (2.7.1) and (2.7.3) would be implemented in an iterative fashion in an actual application. Such an implementation and a simulation are given in this section. The storage requirements and execution time required to simulate the decision rule are discussed.

Equation (2.7.3) can be written iteratively as follows:

$$g_{(\ell)}(x^k,t^k|y_\ell^{n_\ell},\tau_\ell^{n_\ell}) = V_{x_k}^{(\ell)} + K_{x_k}(x^k)\; \frac{\left[W_{x_k}^{(\ell)}+Z_{x_k}(t^{k-1})\right]^{V_{x_k}^{(\ell)}+K_{x_k}(x^{k-1})+1}}{\left[W_{x_k}^{(\ell)}+Z_{x_k}(t^{k})\right]^{V_{x_k}^{(\ell)}+K_{x_k}(x^{k})+1}}$$

$$\times\; \frac{m_{x_{k-1},x_k}^{(\ell)}+N_{x_{k-1},x_k}(x^{k-1})}{m_{x_{k-1}}^{(\ell)}+N_{x_{k-1}}(x^{k-1})} \cdot g_\ell(x^{k-1},t^{k-1}|y_\ell^{n_\ell},\tau_\ell^{n_\ell}) \qquad (2.8.1)$$

$$g(x^1,t^1|y_\ell^{n_\ell},\tau_\ell^{n_\ell}) = (V_{x_1}^{(\ell)}+1)\; \frac{\left(W_{x_1}^{(\ell)}\right)^{V_{x_1}^{(\ell)}+1}}{\left(W_{x_1}^{(\ell)}+Z_{x_1}\right)^{V_{x_1}^{(\ell)}+K_{x_1}+1}}\; P^o(x_1)$$

In (2.8.1), $K_{x_k}(x^k)$ is the number of occupancies of state x_k in x^k, $Z_{x_k}(t^k)$ is the total sojourn time in state x_k in x^k, $N_{x_{k-1},x_k}(x^{k-1})$ is the number of one-step transitions from state x_{k-1} to state x_k in x^{k-1} and $N_{x_{k-1}}(x^{k-1})$ is the number of one-step transitions in x^{k-1} whose initial state is x_{k-1}. An algorithm for computing (2.8.1) is shown in Fig. D.1 of Appendix D.

The optimum-adaptive decision rule implemented in this fashion is a fixed memory rule. No matter how many learning samples are employed and no matter how large the size of the vector to be classified, only the parameters for the posterior densities need be stored. Including the storage requirements for prior information, the total number of words of storage needed is equal to $M(N^2 + 2N + 2)$.

A computer simulation was made to illustrate the performance of the optimum and optimum-adaptive decision rules. All computer simulations discussed below were performed on the CDC 6500 digital computer at Michigan State University. The specific case considered is the following:

1. There are two pattern classes denoted by ω_1 and ω_2; $P_1^o = P_2^o = 1/2$.
2. The continuous-parameter Markov chains which produce the samples are assumed to have 3 states for both pattern classes.
3. The Q-matrices used to produce all observations are listed below.

$$Q_1 = \begin{bmatrix} 0.60 & 0.12 & 0.48 \\ 0.40 & 1.00 & 0.60 \\ 0.36 & 0.84 & 1.20 \end{bmatrix} \qquad Q_2 = \begin{bmatrix} 0.50 & 0.15 & 0.35 \\ 0.60 & 1.20 & 0.60 \\ 0.70 & 0.30 & 1.00 \end{bmatrix}$$

4. The initial state probability distributions for the chains were chosen to be equal to the stationary probability distributions of the corresponding chains. They were computed by (A.11) and are given as follows:

i	$P_1^o(x_1 = i)$	$P_2^o(x_1 = i)$
1	0.387	0.388
2	0.306	0.224
3	0.307	0.388

5. The following parameter matrices were used for the matrix-beta prior density, which was used for the parameters of the jump chain.

$$\mu_1^o = \begin{bmatrix} 0 & 2 & 8 \\ 4 & 0 & 6 \\ 3 & 7 & 0 \end{bmatrix} \qquad \mu_2^o = \begin{bmatrix} 0 & 3 & 7 \\ 5 & 0 & 5 \\ 7 & 3 & 0 \end{bmatrix}$$

6. The following parameters for the prior densities of the sojourn times were used.

$$\underline{v}_1 = \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix}, \quad \underline{w}_1 = \begin{bmatrix} 1.0 \\ 1.0 \\ 1.0 \end{bmatrix} \qquad \underline{v}_2 = \begin{bmatrix} 15 \\ 10 \\ 12 \end{bmatrix}, \quad \underline{w}_2 = \begin{bmatrix} 6.0 \\ 10.0 \\ 12.0 \end{bmatrix}$$

In the first part of the simulation, the average error curve was obtained using the decision rule in (2.3.10). In the second part, the average curves were obtained for the case when the Q-matrices were not known. Two cases of supervised learning were employed, all applying (2.8.1) and differing in the number of training samples provided per pattern class. Training samples of size 50 and 100 were used for each pattern class. Each error curve provides a Monte-Carlo estimate of the probability of error as a function of the size of the observations. In each situation studied, 10 such error curves were obtained along with an average error curve. Using the same notation as [D-4], $100\ell_i(k)$ is defined as the number of wrong decisions in 100 classifications of a sequence (x^k,t^k) for the $i^{\underline{th}}$ run of Fig. D.1. The $k^{\underline{th}}$ point on the average curve is defined by

$$\bar{\ell}(k) = \frac{1}{10} \sum_{i=1}^{10} \ell_i(k)$$

Thus, $100\bar{\ell}(k)$ is the percent of wrong decisions in 1000 independent classifications of k states. Average error curves for several situations are shown in Fig. 2.8.1.

As mentioned in [D-4], the quality and amount of prior information are critical factors in determining the rate at which the error converges to that for known parameters. Equation (2.7.3) shows that the posterior density function for (x^k,t^k), given the training samples, depends on the parameters $V_i^{(\ell)}$, $W_i^{(\ell)}$ and $m_{ij}^{(\ell)}$ where $V_i^{(\ell)} = v_i^{(\ell)} + k_i$, $W_i^{(\ell)} = w_i^{(\ell)} + z_i$, $m_{ij}^{(\ell)} = \mu_{ij}^{o(\ell)} + n_{ij}$. The prior information as presented by $v_i^{(\ell)}$, $w_i^{(\ell)}$ and $\mu_{ij}^{o(\ell)}$, can thus be made to either overwhelm the initial training samples or to be overwhelmed by them so the magnitudes of $v_i^{(\ell)}$, $w_i^{(\ell)}$ and $\mu_{ij}^{o(\ell)}$ are a measure of the amount of prior information being inserted in the decision rule. If these parameters are properly selected, the training data will reinforce the prior information

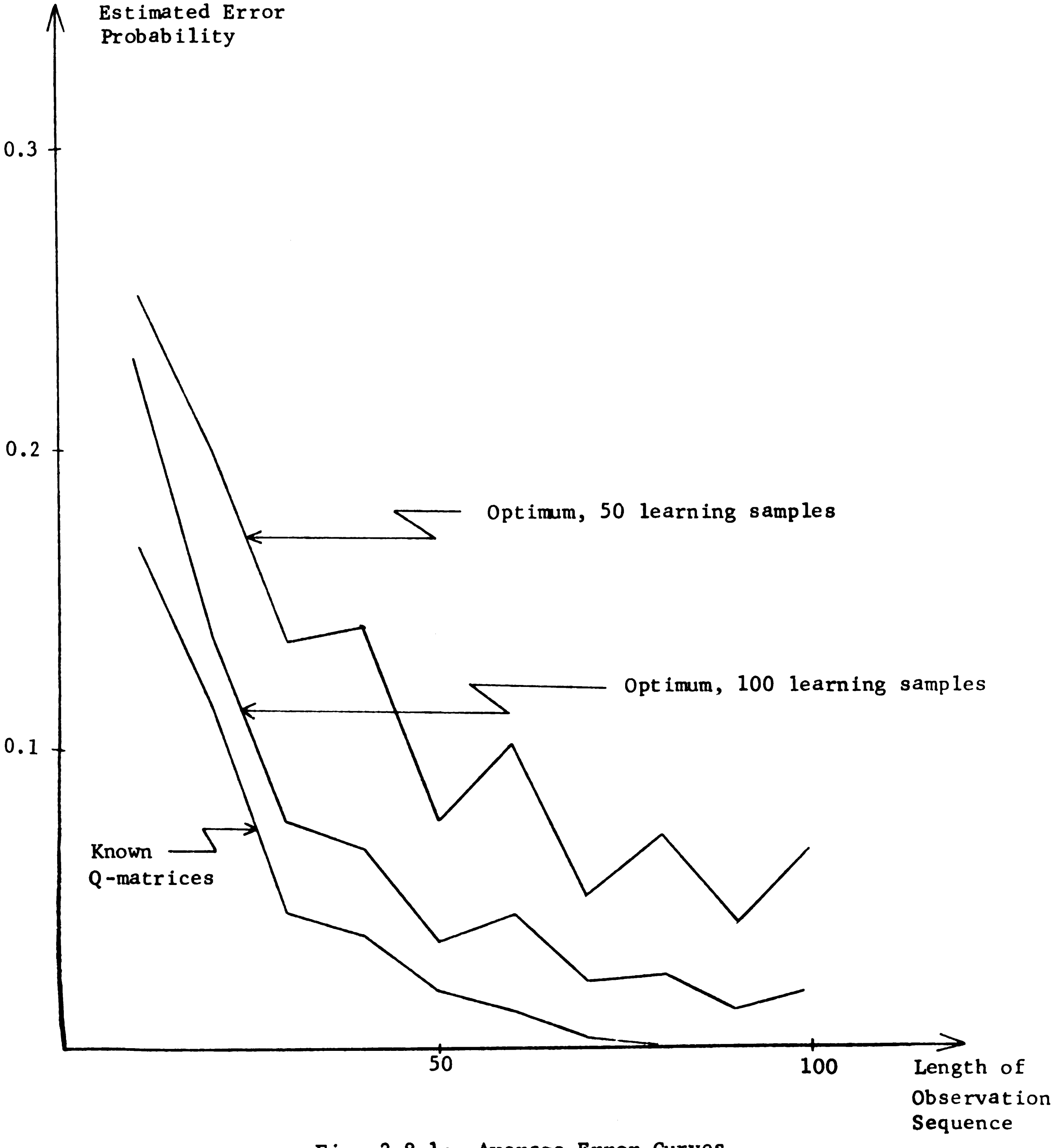


Fig. 2.8.1: Average Error Curves

and the amount of training data used is not critical.

2.9 CONCLUSIONS

This chapter has dealt mainly with decision making and with learning the unknown parameters in finite-state, continuous-parameter Markov system with observable states and transition times. The model by which observations are generated was defined in Sec. 2.1 and the properties of the observation process were given in Sec. 2.2. Assuming the Q-matrices defining the chains were known, an optimal decision rule was defined in Sec. 2.3 to be that rule in a given class of rules which minimizes the Bayes Risk. (B.2).

The optimum-adaptive decision rule (2.4.1) was derived in Sec. 2.4, in case of unknown Q-matrices while the supervised learning scheme for learning them was employed in Sec. 2.5. The existence of a reproducing prior distribution for (q,r) was exhibited. It was also shown that, under the stated conditions, the parameter posterior density (2.5.6) converges to a delta function centered at the true value of the unknown parameter. The posterior densities for the infinitesimal parameters given the training samples were obtained in Sec. 2.6.

The final analytical form of the optimum-adaptive decision rule was given in Sec. 2.7 and it was expressed in an iterative form in Sec. 2.8. Finally, a computer simulation was performed for a specific case to obtain the probability of error curves in both known and unknown parameter cases.

CHAPTER III

PROBABILITY OF ERROR

The quality of a decision rule is characterized by the total probability of error. Unfortunately, for general decision-making problems, exact analytic solutions for the probability of error are impossible. Even if one could find such solutions, they would be tremendously complex. For this reason, simple lower and upper bounds, or asymptotic errors or iterative approximations are more useful than exact error probabilities.

In Sec. 3.1, exact probability of error expressions are derived for the two-pattern class case where both classes are described by 2-state, continuous-parameter Markov chains with known Q-matrices. Lower and upper bounds are given in Sec. 3.2. The limit cases are also studied as the number of observations tends to infinity. Asymptotic probability of error formulas are derived for the two-pattern class problem with N-state Markov chains having known Q-matrices in Sec. 3.3. The probability of error is shown to converge to zero w.p.1 as the number of observations tends to infinity. In Sec. 3.4, the conditional probability of error notion is introduced and iterative expressions are established for them. Finally, Sec. 3.5 summarizes the main results of the chapter.

3.1 EXACT PROBABILITY OF ERROR FOR $N = 2$ AND Q-MATRICES

Let Q_1 and Q_2 be known Q-matrices characterizing pattern classes, ω_1 and ω_2, respectively. The following notation will

be used:

$$Q_1 = \begin{bmatrix} -q_1 & q_1 \\ q_2 & -q_2 \end{bmatrix} \quad ; \quad Q_2 = \begin{bmatrix} -p_1 & p_1 \\ p_2 & -p_1 \end{bmatrix} \quad , \; q_1, q_2, p_1, p_2 > 0$$

The density functions, (2.3.8), are evaluated below for the observation sequence $(x^k, t^k) = (x_1, \ldots, x_k, t_1, \ldots, t_k)$. Since the states must alternate, and assuming k is an even integer, $K_i = k/2$ and $r_{ij} = 1 \;\; \forall \; i,j \in \{1,2\}, \; (j \neq i)$. Then

$$g_1(x^k, t^k) = P_1^o(x_1) \prod_{i=1}^{2} q_i^{k/2} \; e^{-Z_i q_i} \tag{3.1.1}$$

$$g_2(x^k, t^k) = P_2^o(x_1) \prod_{i=1}^{2} p_i^{k/2} \; e^{-Z_i p_i} \tag{3.1.2}$$

Hence, the likelihood ratio, $\Lambda(\cdot)$, is defined as

$$\Lambda(Z_1, Z_2, x_1) \triangleq \frac{g_2(x^k, t^k)}{g_1(x^k, t^k)} = \frac{P_2^o(x_1)}{P_1^o(x_1)} \prod_{i=1}^{2} \left(\frac{p_i}{q_i}\right)^{k/2} e^{-Z_i(p_i - q_i)} \tag{3.1.3}$$

Using the "0-1" loss function, the optimal decision rule is given as

$$\begin{aligned} d^*(x^k, t^k) &= \omega_1 \quad \text{if} \quad L < \eta \\ &= \omega_2 \quad \text{if} \quad L > \eta \end{aligned} \tag{3.1.4}$$

where

$$\eta \triangleq \ell n \left(\frac{P_1^o}{P_2^o}\right)^{2/k} + \ell n \frac{q_1 q_2}{p_1 p_2}; \quad L \triangleq \frac{2}{k} \ell n \frac{P_2^o(x_1)}{P_1^o(x_1)} + \frac{2}{k} \sum_{i=1}^{2} (q_i - p_i) Z_i \tag{3.1.5}$$

The total probability of error, P_e, is then given as

$$P_e = P(L > \eta | \omega_1) P_1^o + P(L < \eta | \omega_2) P_2^o \tag{3.1.6}$$

where
$$P(L > \eta|\omega_1) = P(L > \eta|x_1 = 1,\omega_1)P_1^o(1) + P(L > \eta|x_1 = 2,\omega_1)P_1^o(2) \qquad (3.1.7)$$

$$P(L < \eta|\omega_2) = P(L < \eta|x_1 = 1,\omega_2)P_2^o(1) + P(L < \eta|x_1 = 2,\omega_2)P_2^o(2) \qquad (3.1.8)$$

Using (3.1.5), (3.1.7) and (3.1.8), it follows that

$$P(L > \eta|\omega_1) = P(Z > \eta_1|\omega_1)P_1^o(1) + P(Z > \eta_2|\omega_1)P_1^o(2) \qquad (3.1.9)$$

$$P(L < \eta|\omega_2) = P(Z < \eta_1|\omega_2)P_2^o(1) + P(Z < \eta_2|\omega_2)P_2^o(2) \qquad (3.1.10)$$

where

$$Z = \frac{2}{k}\sum_{i=1}^{2}(q_i-p_i)Z_i \ ; \quad \eta_i = \ln\left[\left(\frac{P_1^oP_1^o(i)}{P_2^oP_2^o(i)}\right)^{2/k} \cdot \frac{q_1q_2}{p_1p_2}\right], \ i=1,2 \qquad (3.1.11)$$

The density functions for Z_1 and Z_2 are now obtained

$$Z_1 = \sum_{i=1}^{k/2} t_{2i-1}; \quad Z_2 = \sum_{i=1}^{k/2} t_{2i} \quad \text{if} \quad x_1 = 1 \qquad (3.1.12)$$

$$Z_1 = \sum_{i=1}^{k/2} t_{2i} \ ; \quad Z_2 = \sum_{i=1}^{k/2} t_{2i-1} \quad \text{if} \quad x_1 = 2 \qquad (3.1.13)$$

There are $k/2$ terms in each sum in the above expressions and also

$t_i \sim$ Exponential (q_i) when ω_1 active

$t_i \sim$ Exponential (p_i) when ω_2 active

Thus, for both values of x_1, Z_i is the sum of $k/2$ i.i.d. random variables, all having the exponential distribution with parameter q_i or p_i $(i = 1,2)$. Thus,

$Z_i \sim$ Gamma $(k/2,q_i)$ under ω_1

$Z_i \sim$ Gamma $(k/2,p_i)$ under ω_2

Explicit probability of error expressions are now derived for several cases.

CASE I. $q_1 > p_1$, $q_2 > p_2$

Let $\xi_1 \triangleq \frac{2}{k}(q_1 - p_1)$; $\xi_2 \triangleq \frac{2}{k}(q_2 - p_2)$; $\xi_1 > 0$, $\xi_2 > 0$. Then, Z, defined in (3.1.11) can be written as

$$Z = \xi_1 Z_1 + \xi_2 Z_2$$

and, $\xi_1 Z_1 \sim$ Gamma $(k/2, \lambda_1)$; $\xi_2 Z_2$ Gamma $(k/2, \lambda_2)$ under ω_1

$\xi_1 Z_1 \sim$ Gamma $(k/2, \mu_1)$; $\xi_2 Z_2$ Gamma $(k/2, \mu_2)$ under ω_2

where $\lambda_i \triangleq \frac{k}{2} \cdot \frac{q_i}{q_i - p_i}$; $\mu_i \triangleq \frac{k}{2} \cdot \frac{p_i}{q_i - p_i}$ $i = 1,2.$ (3.1.14)

The density function for Z can be found by convolving the densities and expanding the integrand with a power series. As a result, the densities for Z under ω_1 and under ω_2 are given as

$$f_Z(z|\omega_1) = \sum_{i=0}^{\infty} c_{i1} z^n e^{-\lambda_2 z} \qquad z > 0 \qquad (3.1.15)$$

$$f_Z(z|\omega_2) = \sum_{i=0}^{\infty} c_{i2} z^n e^{-\mu_2 z} \qquad z > 0 \qquad (3.1.16)$$

where $n \triangleq i + k - 1$; $c_{i1} \triangleq \frac{\Gamma(k/2 + i)}{\Gamma(k+i)\Gamma(k/2)} (\lambda_1 \lambda_2)^{k/2} \frac{(\lambda_2 - \lambda_1)^i}{i!}$;

$$c_{i2} \triangleq \frac{\Gamma(k/2 + i)}{\Gamma(k+i)\Gamma(k/2)} (\mu_1 \mu_2)^{k/2} \frac{(\mu_2 - \mu_1)^i}{i!}$$

Using (3.1.6), (3.1.9), (3.1.10), (3.1.15) and (3.1.16), the total probability of error can be evaluated for several sub-cases.

The following probabilities in (3.1.9) and (3.1.10) are first computed. With a modest amount of effort, it can be shown that

$$P(Z > \eta_1|\omega_1) = \sum_{i=0}^{\infty} b_{i1}\, e^{-\eta_1\lambda_2} \sum_{j=0}^{n} \frac{(\eta_1\lambda_2)^j}{j!} \tag{3.1.17}$$

$$P(Z > \eta_2|\omega_1) = \sum_{i=0}^{\infty} b_{i1}\, e^{-\eta_2\lambda_2} \sum_{j=0}^{n} \frac{(\eta_2\lambda_2)^j}{j!} \tag{3.1.18}$$

$$P(Z < \eta_1|\omega_2) = \sum_{i=0}^{\infty} b_{i2} \left[1 - e^{-\eta_2\lambda_2} \sum_{j=0}^{n} \frac{(\eta_1\lambda_2)^j}{j!}\right] \tag{3.1.19}$$

$$P(Z < \eta_2|\omega_2) = \sum_{i=0}^{\infty} b_{i2} \left[1 - e^{-\eta_2\lambda_2} \sum_{j=0}^{n} \frac{(\eta_2\lambda_2)^j}{j!}\right] \tag{3.1.20}$$

where $b_{i1} \triangleq c_{i1} \dfrac{\Gamma(i+k)}{\lambda_2^{i+k}}$; $b_{i2} \triangleq \dfrac{\Gamma(i+k)}{\mu_2^{i+k}}$

Then, in terms of the above expressions, the total probability of error becomes

$$P_e = P_1^o\left[\sum_{i=0}^{\infty} b_{i1}\left(P_1^o(1)e^{-\eta_1\lambda_2} \sum_{j=0}^{n} \frac{(\eta_1\lambda_2)^j}{j!} + P_1^o(2)e^{-\eta_2\lambda_2} \sum_{j=0}^{n} \frac{(\eta_2\lambda_2)^j}{j!}\right)\right]$$

$$+ P_2^o\left\{\sum_{i=0}^{\infty} b_{i2}\left[P_2^o(1)\left(1 - e^{\eta_1\mu_2} \sum_{j=0}^{n} \frac{(\eta_1\mu_2)^j}{j!}\right) + P_2^o(2)\left(1 - e^{-\eta_2\mu_2} \sum_{j=0}^{n} \frac{(\eta_2\mu_2)^j}{j!}\right)\right]\right\} \tag{3.1.21}$$

CASE IB. $\eta_1 < 0,\ \eta_2 > 0$

$$P(Z > \eta_1|\omega_1) = 1 \ ; \ P(Z < \eta_1|\omega_2) = 0$$

The remaining probabilities are computed as above. Using these probabilities in (3.1.6), the total probability of error is:

$$P_e = P_1^o\left[\sum_{i=0}^{\infty} b_{i1}\left(P_1^o(1) + P_1^o(2)\ e^{-\eta_2\lambda_2} \sum_{j=0}^{n} \frac{(\lambda_2\eta_2)^j}{j!}\right)\right]$$

$$+ P_2^o\left[\sum_{i=0}^{\infty} b_{i2}P_2^o(2)\left(1 - e^{\eta_2\mu_2} \sum_{j=0}^{n} \frac{(\eta_2\mu_2)^j}{j!}\right)\right] \tag{3.1.22}$$

CASE IC. $\eta_1 > 0, \eta_2 < 0$

$$P_e = P_1^o\left[\sum_{i=0}^{\infty} b_{i1}\left(P_1^o(1)e^{-\eta_1\lambda_2}\sum_{j=0}^{n}\frac{(\eta_2\lambda_1)^j}{j!} + P_1^o(2)\right)\right]$$

$$+ P_2^o\left[\sum_{i=0}^{\infty} b_{i2}P_2^o(2)\left(1 - e^{\eta_1\mu_2}\sum_{j=0}^{n}\frac{(\eta_1\mu_2)^j}{j}\right)\right] \qquad (3.1.23)$$

CASE ID. $\eta_1 < 0, \quad \eta_2 < 0$

$$P(Z > \eta_1|\omega_1) = 1 \; ; \; P(Z > \eta_2|\omega_1) = 1$$

$$P(Z < \eta_1|\omega_2) = 0 \; ; \; P(Z < \eta_2|\omega_2) = 0$$

Total probability of error is,

$$P_e = P_1^o$$

CASE II. $q_1 > p_1 \; , \; q_2 < p_2$

Define $\xi_1 \triangleq \frac{2}{k}(q_1 - p_1) \; ; \; \xi_2 = \frac{2}{k}(q_2 - p_2)$ (3.1.24)

Then, the logarithm of likelihood ratio in (3.1.5) becomes

$$L = \frac{2}{k}\ln\frac{P_2^o(x_1)}{P_1^o(x_1)} + \xi_1 Z_1 - |\xi_2|Z_2 \qquad (3.1.25)$$

The total probability of error is calculated in terms of the formulas, (3.1.6), (3.1.9) and (3.1.10), where

$$Z \triangleq \xi_1 Z_1 - |\xi_2|Z_2; \; \eta_i \triangleq \ln\left[\left(\frac{P_1^o P_1^o(i)}{P_2^o P_2^o(i)}\right)^{2/k} \cdot \frac{p_1 p_2}{q_1 q_2}\right], \quad i = 1,2. \qquad (3.1.26)$$

Here, $\xi_1 Z_1 \sim$ Gamma $(k/2,\lambda_1)$; $|\xi_2|Z_2$ Gamma $(k/2,\lambda_2)$ under ω_1

$\xi_1 Z_1 \sim$ Gamma $(k/2,\mu_1)$; $|\xi_2|Z_2$ Gamma $(k/2,\mu_2)$ under ω_2

where $\lambda_i \triangleq \frac{q_i}{|\xi_i|} \; ; \; \mu_i \triangleq \frac{p_i}{|\xi_i|}$, $i = 1,2,.$ (3.1.27)

The density functions for Z, under ω_1 and under ω_2, can be determined as in Case I and are given by:

$$f_Z(z|\omega_1) = \sum_{i=0}^{m} \sum_{j=0}^{n} c_{ij}^{(1)} z^{i+j} e^{-z(\lambda_1+\lambda_2)} \qquad z > 0$$

$$= \sum_{i=0}^{m} c_i^{(1)} z^i e^{z\lambda_2} \qquad z < 0 \qquad (3.1.28)$$

$$f_Z(z|\omega_2) = \sum_{i=0}^{m} \sum_{j=0}^{n} c_{ij}^{(2)} z^{i+j} e^{-z(\mu_1+\mu_2)} \qquad z > 0$$

$$= \sum_{i=0}^{m} c_i^{(2)} z^i e^{z\mu_2} \qquad z < 0 \qquad (3.1.29)$$

where $m \triangleq \frac{k}{2} - 1$; $n \triangleq k - i - 2$

$$c_i^{(1)} \triangleq (-1)^i \binom{m}{i} \frac{\Gamma(n-1)}{(\lambda_1+\lambda_2)^{n-1}} \cdot \frac{(\lambda_1\lambda_2)^{k/2}}{\Gamma^2(k/2)} ;$$

$$c_i^{(2)} \triangleq (-1)^i \binom{m}{i} \frac{\Gamma(n-1)}{(\mu_1+\mu_2)^{n-1}} \cdot \frac{(\mu_1\mu_2)^{k/2}}{\Gamma^2(k/2)}$$

$$c_{ij}^{(1)} \triangleq c_i^{(1)} \cdot \frac{(\lambda_1+\lambda_2)^j}{j!} ; \quad c_{ij}^{(2)} \triangleq c_i^{(2)} \frac{(\mu_1+\mu_2)^j}{j!} .$$

The total probability of error will be computed for the following sub-cases.

CASE IIA. $\eta_1 > 0$, $\eta_2 > 0$

The following probabilities are first computed as in Case I.

$$P(Z > \eta_1|\omega_1) = \sum_{i=0}^{m} \sum_{j=0}^{n} \sum_{r=0}^{\ell} b_{ijr}^{(1)} \eta_1^r e^{-\eta_1(\lambda_1+\lambda_2)} \qquad (3.1.30)$$

$$P(Z > \eta_2|\omega_1) = \sum_{i=0}^{m} \sum_{j=0}^{n} \sum_{r=0}^{\ell} b_{ijr}^{(1)} \eta_1^r e^{-\eta_2(\lambda_1+\lambda_2)} \qquad (3.1.31)$$

$$P(Z < \eta_1|\omega_2) = b^{(2)} - \sum_{i=0}^{m} \sum_{j=0}^{n} \sum_{r=0}^{\ell} b_{ijr}^{(2)} \eta_1^r e^{-\eta_1(\mu_1+\mu_2)} \tag{3.1.32}$$

$$P(Z < \eta_2|\omega_2) = b^{(2)} - \sum_{i=0}^{m} \sum_{j=0}^{n} \sum_{r=0}^{\ell} b_{ijr}^{(2)} \eta_2^r e^{-\eta_2(\mu_1+\mu_2)} \tag{3.1.33}$$

where $m \triangleq \frac{k}{2} - 1$; $n \triangleq k - i - 2$; $\ell \triangleq i + j$,

$$b_{ijr}^{(1)} \triangleq c_{ij}^{(1)} \frac{\Gamma(\ell + 1)}{(\lambda_1+\lambda_2)^{\ell+1}} \cdot \frac{(\lambda_1+\lambda_2)^r}{r!} ;$$

$$b_{ijr}^{(2)} \triangleq c_{ij}^{(2)} \frac{\Gamma(\ell + 1)}{(\mu_1+\mu_2)^{\ell+1}} \cdot \frac{(\mu_1 + \mu_2)^r}{r!} ;$$

$$b_2 \triangleq \sum_{i=0}^{m} \left[(-1)^i \, c_i^{(2)} \frac{\Gamma(i + 1)}{\mu_2^{i+1}} + \sum_{j=0}^{n} c_{ij}^{(2)} \frac{\Gamma(\ell + 1)}{(\mu_1+\mu_2)^{\ell+1}} \right] .$$

Substituting the above expressions in (3.1.9) and (3.1.10) and using them in (3.1.6), the total probability of error is obtained.

$$\begin{aligned} P_e = {} & P_1^o \left\{ \sum_{i,j,r} \left[b_{ijr}^{(1)} \; P_1^o(1)\eta_1^r \, e^{-\eta_1(\lambda_1+\lambda_2)} + P_1^o(2)\eta_2^r \, e^{-\eta_2(\lambda_1+\lambda_2)} \right] \right\} \\ & + P_2^o \left\{ b_2 - \sum_{i,j,r} \left[b_{ijr}^{(1)} \; P_2^o(1)\eta_1^r \, e^{-\eta_1(\mu_1+\mu_2)} + P_2^o(2)\eta_2^r e^{-\eta_2(\mu_1+\mu_2)} \right] \right\} \end{aligned} \tag{3.1.34}$$

Using the same procedures as in Case IIA, the total probability of error expressions are derived for the other sub-cases.

CASE IIB. $\eta_1 < 0$, $\eta_2 > 0$

$$\begin{aligned} P_e = {} & P_1^o \left[b_1 P_1^o(1) - \sum_{i,j} P_1^o(1) b_{ir}^{(1)} (-\eta_1)^r e^{\eta_1\lambda_2} + \sum_{i,j,r} P_1^o(2) b_{ijr}^{(1)} \eta_2^r e^{-\eta_2(\lambda_1+\lambda_2)} \right] \\ & + P_2^o \left[\sum_{i,r} P_2^o(1) b_{ir}^{(2)} (-\eta_1)^r e^{\eta_1\mu_2} + b_2 P_2^o(2) - \sum_{i,j,r} P_2^o(2) b_{ijr}^{(2)} \eta_2^r e^{-\eta_2(\mu_1+\mu_2)} \right] \end{aligned} \tag{3.1.35}$$

CASE IIC. $\eta_1 > 0$, $\eta_2 < 0$

$$P_e = P_1^o\left[\sum_{i,j,r} P_1^o(1) b_{ijr}^{(1)} \eta_1^r e^{-\eta_1(\lambda_1+\lambda_2)} + b_1 P_1^o(2) - \sum_{i,r} P_1^o(2) b_{ir}^{(1)} (-\eta_2)^2 e^{\eta_2\lambda_2}\right]$$

$$+ P_2^o\left[b_2 P_2^o(1) - \sum_{i,j,r} P_2^o(1) b_{ijr}^{(2)} \eta_1^r e^{-\eta_1(\mu_1+\mu_2)} + \sum_{i,r} P_2^o(2) b_{ir}^{(2)} (-\eta_2)^r e^{\eta_2\mu_2}\right] . \tag{3.1.36}$$

CASE IID. $\eta_1 < 0$, $\eta_2 < 0$

$$P_e = P_1^o\left\{b_1 - \sum_{i,r} b_{ir}^{(1)}\left[P_1^o(1)(-\eta_1)^r e^{\eta_1\lambda_2} + P_1^o(2)(-\eta_2)^r e^{\eta_2\lambda_2}\right]\right\}$$

$$+ P_2^o\left\{\sum_{i,r} b_{ir}^{(2)}\left[P_2^o(1)(-\eta_1)^r e^{\eta_1\mu_2} + P_2^o(2)(-\eta_2)^r e^{\eta_2\mu_2}\right]\right\} \tag{3.1.37}$$

where $b_{ir}^{(1)} \triangleq (-1)^i c_i^{(1)} \dfrac{\Gamma(i+1)}{\lambda_2^{i+1}} \dfrac{\lambda_2^r}{r!}$; $b_{ir}^{(2)} \triangleq (-1)^i c_i^{(2)} \dfrac{\Gamma(i+1)}{\mu_2^{i+1}}$;

$$b_1 \triangleq \sum_{i=0}^{m}\left[(-1)^i c_i^{(1)} \frac{\Gamma(i+1)}{\lambda_2^{i+1}} + \sum_{j=0}^{n} c_{ij}^{(1)} \frac{\Gamma(\ell+1)}{(\lambda_1+\lambda_2)^{\ell+1}}\right] .$$

3.2 AN UPPER BOUND ON THE PROBABILITY OF ERROR

The exact probability of error formulas derived in Sec. 3.1 are very complicated and the asymptotic behavior is difficult to ascertain. A simple upper bound on the probability of error is needed and will be derived from the following theorem [K-1], for the case considered in Sec. 3.1.

THEOREM 3.2.1. Let P_1^o and P_2^o be the prior probabilities for the pattern classes, ω_1 and ω_2, respectively, and let $g_i(\cdot)$ be the density function for the sequence of state and sojourn time observations, $(x^k,t^k) = (x_1,\ldots,x_k,t_1,\ldots,t_k)$ when pattern class

ω_i is active, i = 1,2. Then, P_e is bounded by the Bhattacharyya distance as follows:

$$\frac{1}{2}\min(P_1^o P_2^o)\rho^2 \le P_e \le (P_1^o P_2^o)^{\frac{1}{2}}\,\rho \tag{3.2.1}$$

where ρ is Bhattacharyya distance and is defined by

$$\rho \triangleq \int_{\Lambda^k x[0,\infty)^k} \sqrt{g_1(x^k,t^k)g_2(x^k,t^k)}\;dx^k dt^k \tag{3.2.2}$$

To apply this bound to the problem being considered, $g_1(\cdot)$ and $g_2(\cdot)$ must first be determined. Since the 2-state continuous-parameter Markov chain is being considered, the values of $(x_2,\dots,x_k)$ are known w.p.1 as soon as x_1 is known. Substituting $r_{x_i,x_{i+1}} \equiv 1$ i = 1,2,...,k-1 in (2.3.7), the desired densities are determined and are given as:

$$g_1(x^k,t^k) = (q_1q_2)^{k/2}\left[\prod_{i=1}^{k/2} e^{-q_1t_{2i-1}}\, e^{-q_2t_{2i}}\,\delta(1-x_{2i-1})\delta(2-x_{2i})\right]P_1^o(1)$$

$$+\;(q_1q_2)^{k/2}\left[\prod_{i=1}^{k/2} e^{-q_2t_{2i-1}}\, e^{q_1t_{si}}\delta(1-x_{2i})\delta(2-x_{2i-1})\right]P_1^o(2) \tag{3.2.3}$$

$$g_2(x^k,t^k) = (p_1p_2)^{k/2}\left[\prod_{i=1}^{k/2} e^{-p_1t_{2i-1}}\, e^{-p_2t_{si}}\delta(1-x_{2i-1})\delta(2-x_{2i})\right]P_2^o(1)$$

$$+\;(p_1p_2)^{k/2}\left[\prod_{i=1}^{k/2} e^{-p_1t_{2i}}\, e^{-p_2t_{2i-1}}\delta(1-x_{2i})\delta(2-x_{2i-1})\quad P_2^o(2)\right] \tag{3.2.4}$$

Here, $\delta(\cdot)$ is the impulse function. Then,

$$\rho = (q_1q_2p_1p_2)^{k/4}\left[P_1^o(1)P_2^o(1)\prod_{i=1}^{k/2}\int_0^\infty e^{-(\frac{q_1+p_1}{2})t_{2i-1}}dt_{2i-1} \times \int_0^\infty e^{-(\frac{q_2+p_2}{2})t_{2i}}dt_{2i}\right.$$

$$\left.+\;P_1^o(2)P_2^o(2)\prod_{i=1}^{k/2}\int_0^\infty e^{-(\frac{q_2+p_2}{2})t_{2i-1}}dt_{2i-1}\int_0^\infty e^{-(\frac{q_1+p_1}{2})t_{2i}}dt_{2i}\right] \tag{3.2.5}$$

Performing the integrals in the above, the result follows.

$$\rho = (q_1 q_2 p_1 p_2)^{k/4} \left\{ P_1^o(1) P_2^o(1) \left[\frac{4}{(q_1+p_1)(q_2+p_2)} \right]^{k/2} \right.$$

$$\left. + P_1^o(1) P_2^o(2) \left[\frac{4}{(q_2+p_2)(q_1+p_1)} \right]^{k/2} \right\} \qquad (3.2.6)$$

Then, the lower and upper bounds for probability of error are given in terms of ρ, P_1^o, P_2^o by (3.2.1).

Since $(P_1^o P_2^o)^{1/2} \leq \frac{1}{2}$ P_1^o, $P_2^o \in [0,1]$, each term in the upper bound approaches zero as k tends to infinity so P_e approaches zero.

3.3 PROBABILITY OF ERROR FOR LARGE SAMPLE SIZES

In this section, asymptotic probability of error expressions are derived in the case of large sample sizes. In Sec. 3.3.1, the 2-pattern class problem with 2-state continuous-time Markov chains is investigated when both Q-matrices are known while results are generalized to N-state case in Sec. 3.3.2.

3.3.1 N = 2, M = 2, AND KNOWN Q-MATRICES

Let the Q-matrices for the two-pattern classes be given as in Sec. 3.1. The formulas necessary to calculate the total probability of error are given in (3.1.6), (3.1.9) and (3.1.10). For large k, Z_1 and Z_2 defined in (3.1.10) are asymptotically normally distributed. The means m_i and variances σ_i^2, $i = 1,2$ are defined for ω_1 and ω_2 as follows:

$$m_i = \frac{k}{2q_i} \; ; \quad \sigma_i^2 = \frac{k}{2q_i^2} \quad \text{under } \omega_1, \quad i = 1,2$$

$$m_i = \frac{k}{2p_i} \; ; \quad \sigma_i^2 = \frac{k}{2p_i^2} \quad \text{under } \omega_2, \quad i = 1,2$$

Then, Z, defined in (3.1.10) is asymptotically normal with mean m_{ω_i} and variance $\sigma^2_{\omega_i}$ defined by

$$m_{\omega_1} = \frac{q_1-p_1}{q_1} + \frac{q_2-p_2}{q_2} \; ; \quad \sigma^2_{\omega_1} = \frac{2}{k}\left[\frac{(q_1-p_1)^2}{q_1^2} + \frac{(q_2-p_2)^2}{q_2^2}\right] \text{under } \omega_1$$

$$m_{\omega_2} = \frac{q_1-p_1}{p_1} + \frac{q_2-p_2}{p_2} \; ; \quad \sigma^2_{\omega_2} = \frac{2}{k}\left[\frac{(q_1-p_1)^2}{p_1^2} + \frac{(q_2-p_2)^2}{p_2^2}\right] \text{under } \omega_2$$

Thus, the total probability of error for large k is given by the following integral formula:

$$P_e^* = P_1^o\left[P_1^o(1)\int_{\eta_1}^{\infty} f_Z^*(z|\omega_1)dz + P_1^o(2)\int_{\eta_2}^{\infty} f^*(z|\omega_1)dz\right]$$
$$+ P_2^o\left[P_2^o(1)\int_{-\infty}^{\eta_2} f_Z^*(z|\omega_1)dz + P_2^o(2)\int_{-\infty}^{\eta_2} f_Z^*(z|\omega_2)dz\right] \qquad (3.3.1)$$

where η_1 and η_2 are as defined in (3.1.10) and $f_Z^*(z|\omega_i)$, $i = 1,2$, is a limiting density function of the random variable Z. Performing the integrations in (3.3.1), the asymptotic probability of error, P_e^* is obtained.

$$P_e^* = \left\{P_1^o \quad P_1^o(1)\left[1 - \Phi\left(\sqrt{k}\,\frac{\eta_1-m_{\omega_1}}{\sigma_1}\right)\right] + P_1^o(2)\left[1 - \Phi\left(\sqrt{k}\,\frac{\eta_2-m_{\omega_1}}{\sigma_1}\right)\right]\right\}$$
$$+ P_2^o\left[P_2^o(2)\Phi\left(\sqrt{k}\,\frac{\eta_1-m_{\omega_2}}{\sigma_2}\right) + P_2^o(2)\Phi\left(\sqrt{k}\,\frac{\eta_2-m_{\omega_2}}{\sigma_2}\right)\right] \cdot \qquad (3.3.2)$$

where $\sigma_1^2 \triangleq 2\left[\frac{(q_1-p_1)^2}{q_1^2} + \frac{(q_2-p_2)^2}{q_2^2}\right]$; $\sigma_2^2 \triangleq \left[\frac{(q_1-p_1)^2}{p_1^2} + \frac{(q_2-p_2)^2}{p_2^2}\right]$;

$$\Phi(x) \triangleq \int_0^x \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt \;.$$

To show that P_e^* decreases toward zero as k increases without bound, let $q_1 > p_1$, $q_2 > p_2$. Then, it is clear from (3.1.9) that

$$\eta_\infty \triangleq \lim_{k\to\infty} \eta_1 = \lim_{k\to\infty} \eta_2 = \ln \frac{q_1 q_2}{p_1 p_2} > 0$$

A little algebra shows that

$$m_{\omega_1} < \eta_\infty < m_{\omega_2}$$

Thus, $$\lim_{k\to\infty} \Phi\left(\sqrt{k}\, \frac{\eta_1 - m_{\omega_2}}{\sigma_1}\right) = \lim_{k\to\infty} \Phi\left(\sqrt{k}\, \frac{\eta_2 - m_{\omega_1}}{\sigma_1}\right) = 1$$

$$\lim_{k\to\infty} \Phi\left(\sqrt{k}\, \frac{\eta_2 - m_{\omega_2}}{\sigma}\right) = \lim_{k=\infty} \Phi\left(\sqrt{k}\, \frac{\eta_2 - m_{\omega_2}}{\sigma}\right) = 0 \; .$$

So, from (3.3.1), it follows that

$$P_e^* \to 0 \quad \text{as} \quad k \to \infty.$$

3.3.2 N > 2, M = 2 AND KNOWN Q-MATRICES

In this section, the asymptotic probability of error is derived for the N-state case, (N > 2), assuming that there are two pattern classes and that the Q-matrices corresponding to the pattern classes are given.

Let the Q-matrices and the density functions for the observation sequence (x^k, t^k) be given as

$$Q_1 = [q_{ij}],\; g_1(x^k,t^k) = \left[\prod_{i=1}^{N} q_i^{K_i} \exp\{-q_i Z_i\}\right] P_1^o(x_1) \prod_{i=1}^{N} \prod_{\substack{j=1\\ j\neq i}}^{N} \left(\frac{q_{ij}}{q_i}\right)^{N_{ij}} \tag{3.3.3}$$

$$Q_2 = [p_{ij}],\; g_2(x^k,t^k) = \left[\prod_{i=1}^{N} p_i^{K_i} \exp\{-p_i Z_i\}\right] P_2^o(x_1) \prod_{i=1}^{N} \prod_{\substack{j=1\\ j\neq i}}^{N} \left(\frac{p_{ij}}{p_i}\right)^{N_{ij}} \tag{3.3.4}$$

for the ω_1 and ω_2 respectively. Hence, the likelihood ratio, $\Lambda(.)$, defined in (3.1.3) becomes

$$\Lambda(\{K_i\}, \{Z_i\}, \{N_{ij}\}, x_1) = \left[\left(\frac{p_i}{q_i}\right)^{K_i} \exp\{-(p_i - q_i)Z_i\}\right]$$

$$\times \frac{P_2^o(x_1)}{P_1^o(x_1)} \prod_{i=1}^{N} \prod_{\substack{j=1\\ j\neq i}}^{N} \left(\frac{p_{ij}}{p_i} \cdot \frac{q_i}{q_{ij}}\right)^{N_{ij}} . \tag{3.3.5}$$

Using the "0-1" loss function, the optimal decision rule is given as in (3.1.4), where

$$\eta \triangleq \ln\left(\frac{P_1^o}{P_2^o}\right)^{2/k} \quad \text{and} \tag{3.3.6}$$

$$L \triangleq \frac{1}{k} \ln \Lambda$$

$$= \frac{1}{k}\left[\sum_{i=1}^{N} K_i \ln\left(\frac{p_i}{q_i}\right) + \sum_{i=1}^{N} (q_i - p_i)Z_i + \sum_{i=1}^{N} \sum_{\substack{j=1\\ j\neq i}}^{N} N_{ij} \ln\left(\frac{p_{ij}}{p_i} \cdot \frac{q_i}{q_{ij}}\right) + \frac{1}{2} \ln \frac{P_2^o(x_1)}{P_1^o(x_1)}\right]$$

For large k, $K_i \cong \sum_{j=1}^{N} N_{ij}$, then,

$$L = \frac{1}{k} \sum_{i=1}^{N} \sum_{\substack{j=1\\ j\neq i}}^{N} N_{ij} \ln \frac{p_{ij}}{q_{ij}} + \frac{1}{k} \sum_{i=1}^{N} Z_i(q_i - p_i) + \frac{1}{k} \ln \frac{P_2^o(x_1)}{P_1^o(x_1)} . \tag{3.3.7}$$

Since k is fixed and $k = \sum_{i=1}^{N} \sum_{\substack{j=1\\ j\neq i}}^{N} N_{ij}$, the N_{ij}'s are stochastically dependent random variables. Bartlett [B-1] proved that the asymptotic distribution of N_{ij}'s are normal with expected values $m_{ij} = kP_i r_{ij}$, where

$$\begin{aligned} r_{ij} &\triangleq \frac{q_{ij}}{q_i} \quad (j \neq i) \quad \text{under } \omega_1 \\ &\triangleq \frac{p_{ij}}{p_i} \quad (j \neq i) \quad \text{under } \omega_2 \end{aligned} \tag{3.3.8}$$

and the P_i's are the stationary probabilities of the corresponding jump chain. Using Bartlett's method, the covariance matrix $V = KV_0$, where $V_0 = \{\mathrm{cov}(n_{ur}, n_{tq})\}_{u,r,t,q=1}^{N}$, can be determined by the expansion of $\ell n\, \mu_1(\underline{\theta})$ up to second degree in θ_{ij}, where $\mu_1(\underline{\theta})$ is the root such that $\mu_1(0) = 1$, of the matrix

$$R(\underline{\theta}) = \left[r_{ij}\, e^{\theta_{ij}} \right]_{i,j=1}^{N} .$$

Then the matrix V_o which is independent of k is given by the quadratic expression $\frac{1}{2}\, \underline{\theta}^T V_o \underline{\theta}$ in which $\underline{\theta}$ stands for the column vector with components $(\underline{\theta}_1, \underline{\theta}_2, \ldots, \underline{\theta}_N)$ where $\underline{\theta}_i = (\theta_{i1}, \theta_{i2}, \ldots, \theta_i N)$. As a result of the above argument, the first term in (3.3.7), $n \triangleq \frac{1}{k} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j\neq i}}^{N} N_{ij}\, \ell n \frac{p_{ij}}{q_{ij}}$, is asymptotically normally distributed according to $N(m_n, \sigma_n^2)$, where

$$m_n = \sum_{i=1}^{N} \sum_{j=1}^{N} P_i r_{ij}\, \ell n \frac{p_{ij}}{q_{ij}} \tag{3.3.9}$$

$$\sigma_n^2 = \frac{\sigma_o^2}{k} \; ; \; \sigma_o^2 \triangleq \sum_{\substack{u,r,t,q \\ u\neq r,t\neq q}}^{N} (\ell n \frac{p_{ur}}{q_{ur}})(\ell n \frac{p_{tq}}{q_{tq}})\, \mathrm{Cov}(n_{ur}, n_{tq}) \tag{3.3.10}$$

To determine the distribution of $z \triangleq \frac{1}{k} \sum_{i=1}^{N} Z_i(q_i - p_i)$, the second term in (3.3.7), the distribution of the Z_i's is first studied. The random variable Z_i was defined by (3.1.12) and (3.1.13) as

$$Z_i \triangleq \sum_{j=1}^{K_i} t_{ij} \qquad i \in \Lambda \tag{3.3.11}$$

Because of the linear relation $k = \sum_{i=1}^{N} K_i$, the K_i's are dependent random variables. Also, the K_i's are asymptotically normally distributed as k tends to infinity. Means and covariance of the K_i's are given by Good [G-1]. They are as follows:

$$E(K_i) = k\, P_i \qquad i \in \Lambda \tag{3.3.12}$$

$$\mathrm{Cov}(K_i, K_j) = k\, \lambda_{ij} \qquad i,j \in \Lambda \tag{3.3.13}$$

Here, the P_i's are the stationary probabilities of the chain and λ_{ij} is defined as

$$\lambda_{ij} \overset{\Delta}{=} \{\Delta_{ij} P_i - P_i P_j + P_i[S(I-S)^{-1}]_{i,j} + P_j[S(I-S)^{-1}]_{j,i}\}. \tag{3.3.14}$$

The matrix S is defined in terms of the equation given by

$$R = \underline{e}\, \underline{P}^T + S$$

where $R = [r_{ij}]$, $\underline{P}^T = [p_1, p_2, \ldots, p_N]$ and $\underline{e}$ corresponds to left and right eigenvectors, respectively; i.e., $\underline{P}^T R = \underline{P}^T$ and $R\, \underline{e} = \underline{e}$. Now, define

$$y_i \overset{\Delta}{=} \frac{1}{k} \sum_{j=1}^{K_i} t_{ij}(q_i - p_i) \qquad \forall\, i \in \Lambda \tag{3.3.15}$$

Since the K_i's are dependent random variables, the y_i's are also dependent random variables. For small k, the distributions of the y_i's are too involved to compute. However, it can be shown that as k tends to infinity, their distributions approach the normal with probability one (w.p.1). Note that

$$P\{\lim_{k\to\infty} \frac{K_i}{k} = P_i\} \quad \text{w.p.1.} \tag{3.3.16}$$

So, for large k, $K_i \approx k\,P_i$ w.p.1. Substituting this into (3.3.15), it follows that

$$y_i = \frac{1}{k} \sum_{i=1}^{kP_i} t_{ij}(q_i - p_i) \qquad (3.3.17)$$

Since the t_{ij}'s are i.i.d., exponentially distributed random variables, when k is large enough, using the Central limit theorem, it can be seen that the y_i's are independent and asymptotically normal according to $N(m_1, \sigma_i^2)$ where

$$m_i = \frac{P_i(q_i - p_i)}{q_i} \; ; \quad \sigma_i^2 = \frac{P_i(q_i - p_i)^2}{kq_i^2} \quad \text{under } \omega_1$$

$$m_i = \frac{P_i(q_i - p_i)}{p_i} \; ; \quad \sigma_i^2 = \frac{P_i(q_i - p_i)^2}{kq_i^2} \quad \text{under } \omega_2$$

Also, $z = \sum_{i=1}^{N} y_i$ is asymptotically normally distributed according to $N(m_{z_1}, \sigma_{z_1}^2)$ under ω_1 and $N(m_{z_2}, \sigma_{z_2}^2)$ under ω_2, where

$$m_{z_1} = \sum_{i=1}^{N} \frac{P_i(q_i - p_i)}{q_i} \; ; \quad \sigma_{z_1}^2 = \frac{1}{k} \sum_{i=1}^{N} \frac{P_i(q_i - p_i)^2}{q_i^2} \qquad (3.3.18)$$

$$m_{z_2} = \sum_{i=1}^{N} \frac{P_i(q_i - p_i)}{p_i} \; ; \quad \sigma_{z_2}^2 = \frac{1}{k} \sum_{i=1}^{N} \frac{P_i(q_i - p_i)}{p_i^2} \, . \qquad (3.3.19)$$

As a result, $Z \overset{\Delta}{=} n + z$ is asymptotically normally distributed according to $N(m_{Z_i}, \sigma_{Z_i}^2)$, $i = 1,2$. Where

$$m_{Z_1} = m_{z_1} + m_{n_1} \; ; \quad \sigma_{Z_1}^2 = \sigma_{z_1}^2 + \sigma_{n_1}^2 \quad \text{under } \omega_1 \qquad (3.3.20)$$

$$m_{Z_2} = m_{z_2} + m_{n_2} \; ; \quad \sigma_{Z_2}^2 = \sigma_{z_2}^2 + \sigma_{n_2}^2 \quad \text{under } \omega_2 . \qquad (3.3.21)$$

In the above, m_{n_i} and $\sigma^2_{n_i}$, $i = 1,2$, are expressed in terms of m_n and σ^2_n defined in (3.3.9) and (3.3.10) as follows:

$$m_{n_1} = m_n \; ; \; \sigma^2_{n_1} = \sigma^2_n \quad \text{when} \quad r_{ij} = \frac{q_{ij}}{q_i} \quad i,j \in \Lambda \quad (j \neq i)$$

$$m_{n_2} = m_n \; ; \; \sigma^2_{n_2} = \sigma^2_n \quad \text{when} \quad r_{ij} = \frac{p_{ij}}{p_i} \quad i,j \in \Lambda \quad (j \neq i)$$

The total probability of error, then, can be calculated using equations (3.16), (3.19) and (3.1.10). The result is as follows:

$$P^*_e = P^o_1 \; P_1(1) \; 1 - \Phi\left(\sqrt{k}\,\frac{\eta_1 - m_{Z_1}}{\sigma^*_{Z_1}}\right) + P_1(2) \;\; 1 - \Phi\left(\sqrt{k}\,\frac{\eta_2 - m_{Z_1}}{\sigma^*_{Z_1}}\right)$$

$$+ P^o_2 \; P_2(1) \; \Phi\left(\sqrt{k}\,\frac{\eta_1 - m_{Z_2}}{\sigma^*_{Z_2}}\right) + P_2(2) \; \Phi\left(\sqrt{k}\,\frac{\eta_2 - m_{Z_2}}{\sigma^*_{Z_2}}\right) \qquad (3.3.22)$$

where $\sigma^*_{Z_i} \triangleq \sqrt{k}\,\sigma_{Z_i}$, $\eta_i \triangleq \ell n \left[\frac{P^o_1}{P^o_2}\,\frac{P^o_1(i)}{P^o_2(i)}\right]^{2/k}$, $i = 1,2.$

Note that $\lim_{k \to \infty} \eta_i = 0$, $i = 1,2.$

It is shown that for a specific case in which $q_{ij} > p_{ij}$, $\forall$ $i,j \in \Lambda$ $(j \neq i)$, the total probability of error decreases toward zero as k tends to infinity. The following lemma is first introduced and its proof is given in Appendix E.

LEMMA 3.3.2 $\quad q_{ij} > p_{ij} \Rightarrow -\infty < m_{Z_1} < 0 < m_{Z_2} < +\infty$

From Lemma 3.3.2 and the property of $\Phi(.)$, the following limits hold:

$$\lim_{k\to\infty} \Phi\left(\sqrt{k}\,\frac{\eta_1 - m_{z_1}}{\sigma^*_{z_1}}\right) = \lim_{k\to\infty} \Phi\left(\sqrt{k}\,\frac{\eta_2 - m_{z_1}}{\sigma^*_{z_1}}\right) = 1$$

$$\lim_{k\to\infty} \Phi\left(\sqrt{k}\,\frac{\eta_1 - m_{z_2}}{\sigma^*_{z_2}}\right) = \lim_{k\to\infty} \Phi\left(\sqrt{k}\,\frac{\eta_2 - m_{z_2}}{\sigma^*_{z_2}}\right) = 0$$

Thus, from (3.3.20), it follows that

$$\lim_{k\to\infty} P^*_e = 0.$$

3.4 RECURRENT EXPRESSIONS FOR THE PROBABILITY OF ERROR

In two-pattern class problems, the total probability of error can always be expressed in terms of the probability of error of the first kind (the probability of false alarm), α_o, the second kind (the probability of a false dismissal), β_om and the prior probabilities, P^o_1, P^o_2.

$$P_e = P^o_1\alpha_o + P^o_2\beta_o \qquad (3.4.1)$$

where α_o and β_o are defined as follows:

$$\alpha_o \triangleq P[d^*(x^k,t^k) = \omega_2 | \theta = 1],\ \beta_o \triangleq P[d^*(x^k,t^k) = \omega_1 | \theta = 2] \qquad (3.4.2)$$

From (3.1.4), they can be written in another form as

$$\alpha_o \triangleq P[L_K > \eta | \theta = 0],\ \beta_o \triangleq P[L_K < \eta | \theta = 1] \qquad (3.4.3)$$

where L_K is the logarithm of the likelihood ratio and η is a threshold.

In general, exact analytical expressions for α_o, and β_o are impossible. However, it is possible to obtain them recursively in terms of the conditional error probabilities, $\alpha_\ell(k)$ and $\beta_\ell(k)$, where $\alpha_\ell(k)$ is the probability of $d^*(x^k,t^k) = \omega_2$, given that $\theta = 1$ and $x^\ell = (x_1,\ldots,x_\ell)$, $\ell \le k$ while $\beta_\ell(k)$ is the probability of $d^*(x^k,t^k) = \omega_1$, given that $\theta = 2$ and $x^\ell = (x_1,\ldots,x_\ell)$. For short, this can be written as,

$$\alpha_\ell(k) = P[d^*(x^k,t^k) = \omega_2 | \theta = 1, x^\ell] = P[L_K > \eta | \theta = 1, x^\ell] \tag{3.4.4}$$

$$\beta_\ell(k) = P[d^*(x^k,t^k) = \omega_1 | \theta = 2, x^\ell] = P[L_K < \eta | \theta = 2, x^\ell] \tag{3.4.5}$$

It is clear that the usual error probabilities, α_o, β_o are related to the conditional error probabilities by the following expressions:

$$\alpha_o = \alpha_o(k) \ , \quad \beta_o = \beta_o(k) \tag{3.4.6}$$

With the help of the conditional error probabilities and the total probability law, α_o and β_o may be obtained in terms of the following expressions:

$$\alpha_o = \sum_{x^\ell} \int_{t^\ell} \alpha_\ell(k) g_1(x^\ell,t^\ell) dt^\ell \tag{3.4.7}$$

$$\beta_o = \sum_{x^\ell} \int_{t^\ell} \beta_\ell(k) g_2(x^\ell,t^\ell) dt^\ell \tag{3.4.8}$$

where $x^\ell \in \Lambda^\ell$, $t^\ell \in [0,\infty)^\ell$, $\ell = 0,1,2,\ldots,k$.

The conditional probabilities of error can be generated iteratively. This can be achieved writing $g_i(x^\ell,t^\ell)$, $i = 1,2$, as:

$$g_i(x^\ell,t^\ell) = f_i(x_\ell | x^{\ell-1},t^{\ell-1}) f_i(t_\ell | x^\ell,t^{\ell-1}) g_i(x^{\ell-1},t^{\ell-1}) \tag{3.4.9}$$

Substituting in (3.4.7):

$$\alpha_o = \sum_{x^{\ell-1}} \int_{t^{\ell-1}} g_1(x^{\ell-1},t^{\ell-1}) \sum_{x^{\ell}} \int_{t^{\ell}} \alpha_{\ell}(k) f_1(x_{\ell}|x^{\ell-1},t^{\ell-1}) f_1(t_{\ell}|x^{\ell},t^{\ell-1}) dt_{\ell}\, dt^{\ell-1}$$

Defining

$$\alpha_{\ell-1}(k) \overset{\Delta}{=} \sum_{x^{\ell}} \int_{t^{\ell}} \alpha_{\ell}(k) f_1(x_{\ell}|x^{\ell-1},t^{\ell-1}) f_1(t_{\ell}|x^{\ell},t^{\ell-1}) dt_{\ell},\ \ell = 1,\ldots,k \tag{3.4.10}$$

the desired iterative form for $\alpha_{\ell}(k)$ is obtained. Similarly, for $\beta_{\ell}(k)$:

$$\beta_{\ell-1}(k) = \sum_{x^{\ell}} \int_{t^{\ell}} \beta_{\ell}(k) f_2(x_{\ell}|x^{\ell-1},t^{\ell-1}) f_2(t_{\ell}|x^{\ell},t^{\ell-1}) dt_{\ell},\ \ell = 1,\ldots,k. \tag{3.4.11}$$

From (3.4.4) and (3.4.5), it follows that these probabilities of error must satisfy the following boundary conditions when $\ell = k$.

$$\alpha_k(k) = u(L_k - \eta), \quad \beta_k(k) = 1 - u(L_k - \eta) \tag{3.4.12}$$

Here, u(.) is the unit step function. In terms of the recurrent relations (3.4.10), (3.4.11) and of the boundary conditions, (3.4.12), it is possible to obtain P_e via α_o and β_o. Unfortunately, except for some simple cases, an analytical form for the total probability of error is almost impossible to obtain. However, because of the iterative character of the method, it is more convenient for the computer implementation than the work in Sec. 3.3.

3.5 CONCLUSIONS

The aim of this chapter has been to derive expressions for the total probability of error and to study their asymptotic properties for the problem considered in Chapter II. Because of the tremendous complexity involved in computations with unknown parameters, attempts have been made only for the known parameters case with two pattern classes.

The exact analytical solutions were obtained in Sec. 3.1 for the total probability of error assuming two pattern classes with 2-state chains and known Q-matrices while lower and upper bounds were given in Sec. 3.2. It was shown that the probability of error decreases toward zero as the number of observations increases without bound. Section 3.3 derives the asymptotic probability of errors for large sample sizes. Proving the asymptotic normalities of the underlying distributions, asymptotic error expressions were obtained for the case of two pattern classes, N-states and known Q-matrices, and, in terms of the Lemma 3.3.2, their convergence to zero as the number of observations tends to infinity was proven.

Finally, In Sec. 3.4, the conditional probabilities of error of the first and second kind were defined and it was shown that the total probability of error can be computed in terms of these conditional error probabilities and of the prior probabilities. A method was also given to generate them in an iterative fashion.

CHAPTER IV

DECISION MAKING AND LEARNING WITH UNOBSERVABLE STATES AND OBSERVABLE TRANSITION TIMES

In this chapter, optimal decision making with unreliable observations on finite-state, continuous-parameter Markov systems is considered. The states of the systems cannot be observed directly, but the exact times at which transitions from one state to another occur can be observed. The decision problem itself is the same as that in Chapter II.

The model by which observations are generated is defined in Sec. 4.1. In Sec. 4.2, the optimal decision rule is established and is generated iteratively for the case when all parameters in the model are assumed known. In Sec. 4.3, the results of Sec. 4.2 are applied to two specific cases. Section 4.4 deals with various aspects of decision making for continuous-parameter Markov systems when the model is not completely specified. The main problem of Sec. 4.4 is to extract some information about the unknowns of the model from the observations and use this information to construct an optimum-adaptive decision rule which performs almost as well (with respect to a well-defined criteria) as the optimal decision rule of Sec. 4.3. Finally, Sec. 4.5 summarizes the main results of the chapter.

4.1 MATHEMATICAL MODEL FOR GENERATING OBSERVATIONS

The basic model considered in this chapter is similar to that in Chapter II. Each of the M-pattern classes is defined by a finite-state, continuous-parameter Markov chain. The Q-matrices for the chains are parameters of the problem. The states of the chains cannot be observed directly, but each state is characterized by an observable random process. For instance, a noise process can be added to each state. The random variables describing the observations during a given time interval have joint distributions which depend on the state of the continuous-parameter Markov chain during that time interval.

Appendix A shows that almost all sample functions of a finite-state, continuous-parameter Markov chain, $\{x_t, 0 \le t < \infty\}$, which satisfies certain conditions are step functions, and that the process can be uniquely determined by observing any of the sample functions in the interval $0 \le t < \infty$. Observing the sample function on $[0,\infty)$ is equivalent to observing the sequences of random variables $\{\lambda_k\}_1^\infty$, $\{t_k\}_1^\infty$ corresponding to the state numbers and the sojourn times. It is also shown in Appendix A that $\{\lambda_k\}_1^\infty$ is a discrete-time Markov chain (jump chain) with the transition matrix $[r_{ij}]$ defined in (2.2.1) and that the sojourn time random variables $\{t_k\}_1^\infty$ are state-conditionally independent and have exponential distributions with parameters $\{q_i\}_1^N$.

In this chapter, the random variables $\{\lambda_k\}_1^\infty$ cannot be observed directly, but it is known that when $\lambda_k = i$, a sample function of a random process $x_i(t)$ can be observed. The random variables $\{t_k\}_1^\infty$ can be observed. The main assumption on the random processes $\{x_i(t)\}_1^N$, which describe the states of the chains generated by each pattern class, is that the statistical properties of the random processes $\{x_i(t)\}_1^N$ are known and are the same for all pattern classes. This simplifies the formulation.

There are many possible sampling schemes for defining the observation process. Here, it is assumed that time samples $x(t_{ij}) \triangleq x_{ij}$, $i = 1,2,\ldots;\ j = 1,2,\ldots,n_i$, are taken during the interval $[\rho_{i-1},\rho_i)$ where ρ_{i-1} and ρ_i are the successive jump points of the process at which transitions occur and $\rho_o \triangleq 0$. Another sampling scheme is discussed in Sec. 4.3.

The observation process is then defined by the sequence of random variables $\{\underline{x}^k,t^k\}_1^\infty$ where $t^k \triangleq (t_1,\ldots,t_k)$, $t_i \triangleq \rho_i - \rho_{i-1}$, and $\underline{x}^k \triangleq (\underline{x}_1,\underline{x}_2,\ldots,\underline{x}_k)$; $\underline{x}_i$ corresponds to the $i^{\underline{th}}$ observed vector $\underline{x}_i^T = [x_{i1},\ldots,x_{in_i}]$. Sampling times t_{ij} are determined as follows: The first sample x_{i1} is always taken at time $t_{i1} = \rho_{i-1}$, just after a transition occurs. The following samples $(x_{i2},\ldots,x_{in_i})$ are taken at times $t_{i2} = \rho_{i-1} + \tau,\ldots,t_{in_i} = \rho_{i-1} + (n_i-1)\tau$, respectively, where $\tau > 0$ is the time interval between two samples. From the above, it is clear that the number of samples n_i is a random variable which depends on t_i and is determined by $n_i = [t_i/\tau]$, in which the expression $[\xi]$ is defined for any real number $\xi \geq 0$ as the largest integer less than or equal to ξ. The general model with the sampling scheme defined above is illustrated in Fig. 4.1.1. The random vector $\underline{x}_k^T = [x_{k1},x_{k2},\ldots,x_{kn_k}]$ defined above takes values in R^{n_k} and has a n_k-dimensional density function denoted by $f_i(.|\theta)$ when it is known that pattern class θ is active and that the corresponding Markov chain is in state i

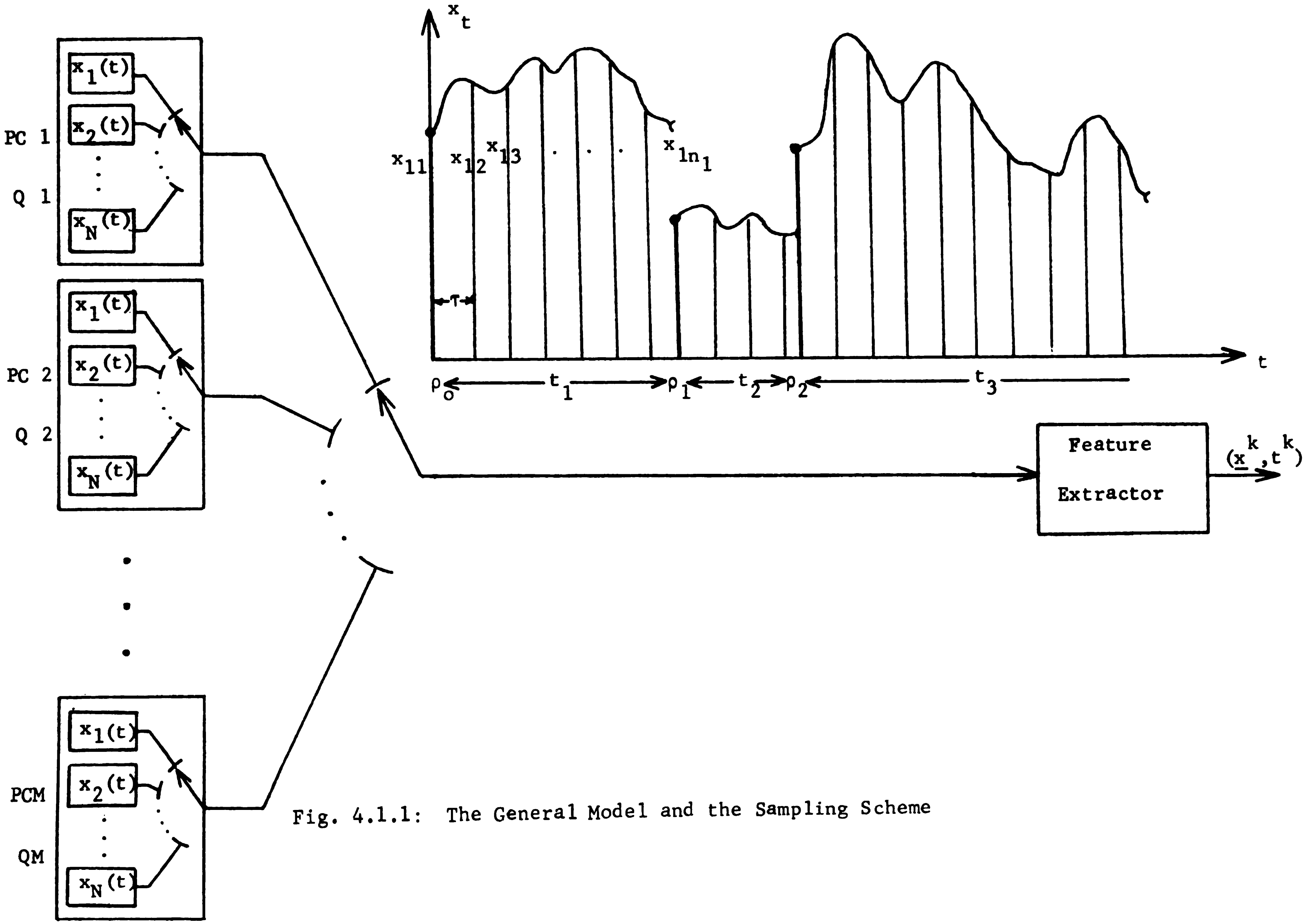


Fig. 4.1.1: The General Model and the Sampling Scheme

during the interval $[\rho_{i-1},\rho_i)$. That is, $f_i(\cdot|\theta)$ is the conditional density of observations $\underline{x}_k$ given that chain θ is in state i. By assumption, $f_i(\underline{x}_k|\theta)$ is independent of θ for any $\underline{x}_k$ and any k and will be denoted by $f_i(\underline{x}_k)$. Since the states of the underlying continuous-parameter Markov chains are not observed, $\underline{x}_k$ has the following global density when pattern class θ is active:

$$f(\underline{x}_k|\theta) = \sum_{i=1}^{N} f_i(\underline{x}_k)P(\lambda_k = i|\theta) \qquad \theta \in \Theta \tag{4.1.1}$$

This density is a finite-mixture with component densities $\{f_i(\cdot)\}_1^N$ and mixing parameters $\{P(\lambda_k = i|\theta)\}_1^N$ where $\sum_{i=1}^{N} P(\lambda_k = i|\theta) = 1$. Another simplifying assumption is that the random variables in the sequence $\{\underline{x}^k\}_1^\infty$ are state-conditionally independent. That is,

$$f_i(\underline{x}_1,\underline{x}_2,\ldots,\underline{x}_k|\lambda_1 = i,\lambda_2 = j,\ldots,\lambda_k = \ell,\theta) = f_i(\underline{x}_1)f_j(\underline{x}_2)\ldots f_\ell(\underline{x}_k) \tag{4.1.2}$$

As in Chapter II, a single, continuous-parameter Markov chain is active to produce $(\underline{x}^k,t^k)$ during the entire observation interval $[0,T]$, $T < \infty$. A decision about the identity of that chain is to be made.

4.2 ITERATIVE GENERATION OF THE OPTIMUM DECISION RULE

The model is completely specified when the following quantities are known: The compoennt densities, $\{f_i(\cdot)\}_1^N$, the infinitesimal parameters, $\{q_{ij}^{(\theta)}\}_1^N$, and the initial probabilities, $\{P(\lambda_1=i|\theta)\}_1^N$ over the states for all pattern classes $\theta \in \Theta$. Decision theory then provides the optimal strategies for processing the observations.

An optimum, non-randomized decision rule $d^*(\underline{x}^k,t^k)$ can be chosen as in Appendix B. In particular, when the loss function is given by $L(i,j) = 1 - \Delta_{ij}$ (0-1 loss function), then the optimum decision rule becomes a minimum probability of error rule defined by

$$d^*(\underline{x}^k,t^k) = s \quad \text{if} \quad P(\theta=s|\underline{x}^k,t^k) = \left[\max_{\ell\in\Theta} \{P(\theta=\ell|\underline{x}^k,t^k)\}\right]_{\ell=s}$$

or

$$\text{if} \quad P_s^o g_s(\underline{x}^k,t^k) = \left[\max_{\ell\in\Theta} \{P_\ell^o g_\ell(\underline{x}^k,t^k)\}\right]_{\ell=s} \qquad (4.2.1)$$

The rule above can be written iteratively. For simplicity, the subscript ℓ denoting pattern class will be dropped. From the Markov property

$$g(\underline{x}^k,t^k) = f(t_k|\underline{x}^k,t^{k-1})f(\underline{x}_k|\underline{x}^{k-1},t^{k-1})g(\underline{x}^{k-1},t^{k-1}) \qquad (4.2.2)$$

The last factor, $g(\underline{x}^{k-1},t^{k-1})$, is available from the previous step of the iteration scheme. The middle factor, $f(\underline{x}_k|\underline{x}^{k-1},t^{k-1})$, can be written in an iterative manner as follows. By the total probability law,

$$f(\underline{x}_k|\underline{x}^{k-1},t^{k-1}) = \sum_{i=1}^{N} f(\underline{x}_k|\lambda_k=i,\underline{x}^{k-1},t^{k-1})P(\lambda_k=i|\underline{x}^{k-1},t^{k-1}) \qquad (4.2.3)$$

where Assumption (4.1.2) implies

$$f(\underline{x}_k|\lambda_k=i,\underline{x}^{k-1},t^{k-1}) = f_i(\underline{x}_k) \qquad (4.2.4)$$

The second factor in (4.2.3) will also be required in the first part of (4.2.2) and is computed below. This factor, $P(\lambda_k=i|\underline{x}^{k-1},t^{k-1})$, can be obtained iteratively as follows:

$$P(\lambda_k=i|\underline{x}^{k-1},t^{k-1}) = \sum_{\substack{j=1\\ j\neq i}}^{N} P(\lambda_k=i|\lambda_{k-1}=j,\underline{x}^{k-1},t^{k-1})P(\lambda_{k-1}=j|\underline{x}^{k-1},t^{k-1}) \tag{4.2.5}$$

From (4.2.1),

$$\begin{aligned} P(\lambda_k=i|\lambda_{k-1}=j,\underline{x}^{k-1},t^{k-1}) &= r_{ji} \quad \text{if} \quad j \neq i \\ &= 0 \quad \text{if} \quad j = i \end{aligned} \tag{4.2.6}$$

Using Bayes rule for the second factor in summation (4.2.5),

$$P(\lambda_{k-1}=j|\underline{x}^{k-1},t^{k-1}) = \frac{f(t_{k-1}|\lambda_{k-1}=j,\underline{x}^{k-1},t^{k-2})P(\lambda_{k-1}=j|\underline{x}^{k-1},t^{k-2})}{\sum_{j=1}^{N} f(t_{k-1}|\lambda_{k-1}=j,\underline{x}^{k-1},t^{k-2})P(\lambda_{k-1}=j|\underline{x}^{k-1},t^{k-2})} \tag{4.2.7}$$

where, from the state-conditionality independence assumption

$$f(t_{k-1}|\lambda_{k-1}=j,\underline{x}^{k-1},t^{k-1}) = f(t_{k-1}|\lambda_{k-1}=j) = q_j \exp\{-q_j t_{k-1}\} \tag{4.2.8}$$

and

$$P(\lambda_{k-1}=j|\underline{x}^{k-1},t^{k-2}) = \frac{f_i(\underline{x}_{k-1})P(\lambda_{k-1}=j|\underline{x}^{k-2},t^{k-2})}{\sum_{m=1}^{N} f_m(\underline{x}_{k-1})P(\lambda_{k-1}=m|\underline{x}^{k-2},t^{k-2})} \tag{4.2.9}$$

Then, putting (4.2.6) and (4.2.7) into (4.2.5),

$$P(\lambda_k=i|\underline{x}^{k-1},t^{k-1}) = \sum_{\substack{j=1\\ j\neq i}}^{N} r_{ji} \frac{f(t_{k-1}|\lambda_{k-1}=j)f_j(\underline{x}_{k-1})P(\lambda_{k-1}=j|\underline{x}^{k-2},t^{k-2})}{\sum_{m=1}^{N} f(t_{k-1}|\lambda_{k-1}=m)f_m(\underline{x}_{k-1})P(\lambda_{k-1}=m|\underline{x}^{k-2},t^{k-2})} \tag{4.2.10}$$

This is a "predictive" factor since it used the present and past samples to make decisions about future states. Substituting this in (4.2.3) will produce the middle factor of (4.2.2).

The first factor in (4.2.2), $f(t_k|\underline{x}^k,t^{k-1})$, can be written in an iterative form as follows: By the total probability law and the Bayes rule,

$$f(t_k|\underline{x}^k,t^{k-1}) = \sum_{i=1}^{N} f(t_k|\lambda_k=i,\underline{x}^k,t^{k-1})P(\lambda_k=i|\underline{x}^k,t^{k-1}) \qquad (4.2.11)$$

where

$$f(t_k|\lambda_k=i,\underline{x}^k,t^{k-1}) = f(t_k|\lambda_k=i) = q_i \exp\{-q_i t_k\} \qquad (4.2.12)$$

$$P(\lambda_k=i|\underline{x}^k,t^{k-1}) = \frac{f_i(\underline{x}_k)P(\lambda_k=i|\underline{x}^{k-1},t^{k-1})}{\sum_{m=1}^{N} f_m(\underline{x}_k)P(\lambda_k=m|\underline{x}^{k=1},t^{k-1})} \qquad (4.2.13)$$

Substituting (4.2.12) and (4.2.13) in (4.2.11),

$$f(t_k|\underline{x}^k,t^{k-1}) = \sum_{i=1}^{N} f(t_k|\lambda_k=i) \frac{f_i(\underline{x}_k)P(\lambda_k=i|\underline{x}^{k-1},t^{k-1})}{\sum_{m=1}^{N} f_m(\underline{x}_k)P(\lambda_k=m|\underline{x}^{k-1},t^{k-1})} \qquad (4.2.14)$$

Here, again, $\{P(\lambda_k=i|\underline{x}^{k-1},t^{k-1})\}_{i=1}^{N}$ can be obtained iteratively from (4.2.2).

The starting procedure for the iterative scheme is given below. When $k = 1$, observe (x_1,t_1) and compute

$$g(\underline{x}^1,t^1) = f(t_1|\underline{x}_1)f(\underline{x}_1) \qquad (4.2.15)$$

where

$$f(\underline{x}_1) = \sum_{i=1}^{N} f_i(x_1)P_i^o , \quad P_i^o \triangleq P(\lambda_1 = i)$$

$$f(t_1|\underline{x}_1) = \sum_{i=1}^{N} f(t_1|\lambda_1=i) \frac{f_i(\underline{x}_1)P_i^o}{\sum_{m=1}^{N} f_m(\underline{x}_1)P_m^o}$$

Next, compute the predictive probabilities, $\{P(\lambda_2=i|x_1,t_1)\}_{i=1}^{N}$.

$$P(\lambda_2=i|\underline{x}^1,t^1) = \sum_{\substack{j=1 \\ j\neq i}}^{N} r_{ji} \frac{f(t_1|\lambda_1=j)f_j(\underline{x}_1)P_j^o}{\sum_{m=1}^{N} f(t_1|\lambda_1=m)f_m(\underline{x}_1)P_m^o} .$$

When $k = 2$, observe (x_2,t_2), compute

$$f(\underline{x}^2,t^2) = f(t_2|\underline{x}^2,t^1)f(\underline{x}_2|\underline{x}^1,t^1)f(\underline{x}^1,t^1)$$

where $f(\underline{x}^1,t^1)$ is obtained from the above, and

$$f(\underline{x}_2|\underline{x}^1,t^1) = \sum^N f_i(\underline{x}_2)P(\lambda_2=i|\underline{x}^1,t^1)$$

$$f(t_2|\underline{x}^2,t^1) = \sum_{i=1}^N f(t_2|\lambda_2=i)\frac{f_i(\underline{x}_2)P(\lambda_2=i|\underline{x}^1,t^1)}{\sum_{m=1}^N f_m(\underline{x}_2)f(\lambda_2=m|\underline{x}^1,t^1)}$$

Again, $\{P(\lambda_3=i|\underline{x}^2,t^2)\}_{i=1}^N$ must be computed for the next step. This procedure can be repeated up to time k. The flow-diagram 4.2.1 shows how $g_s(x^k,t^k)$ can be determined in an iterative fashion.

4.3 APPLICATION OF THE MODEL TO SPECIAL CASES

The model defined in Sec. 4.1 is quite general and applicable to problems encountered in communication theory, pattern recognition and operations research. The major difficulty in implementing the optimal decision rule is that even if all information about the random processes $\{x_i(t)\}_{i=1}^N$ is known, determining the n_i^{th} order density of $\underline{x}_i = (x_{i1},x_{i2},\ldots,x_{in_i})$ can be very complex and, sometimes, impossible. Some ways of getting around this problem are presented below.

1. Take only one sample from each time interval $[\rho_{i-1},\rho_i)$. The random vector $\underline{x}_i$ is then reduced to a single random variable, x_i. Sampling points may be taken at times ρ_{i-1} when transitions occur. Fig. 4.3.2 shows the sampling scheme for a particular sample function.

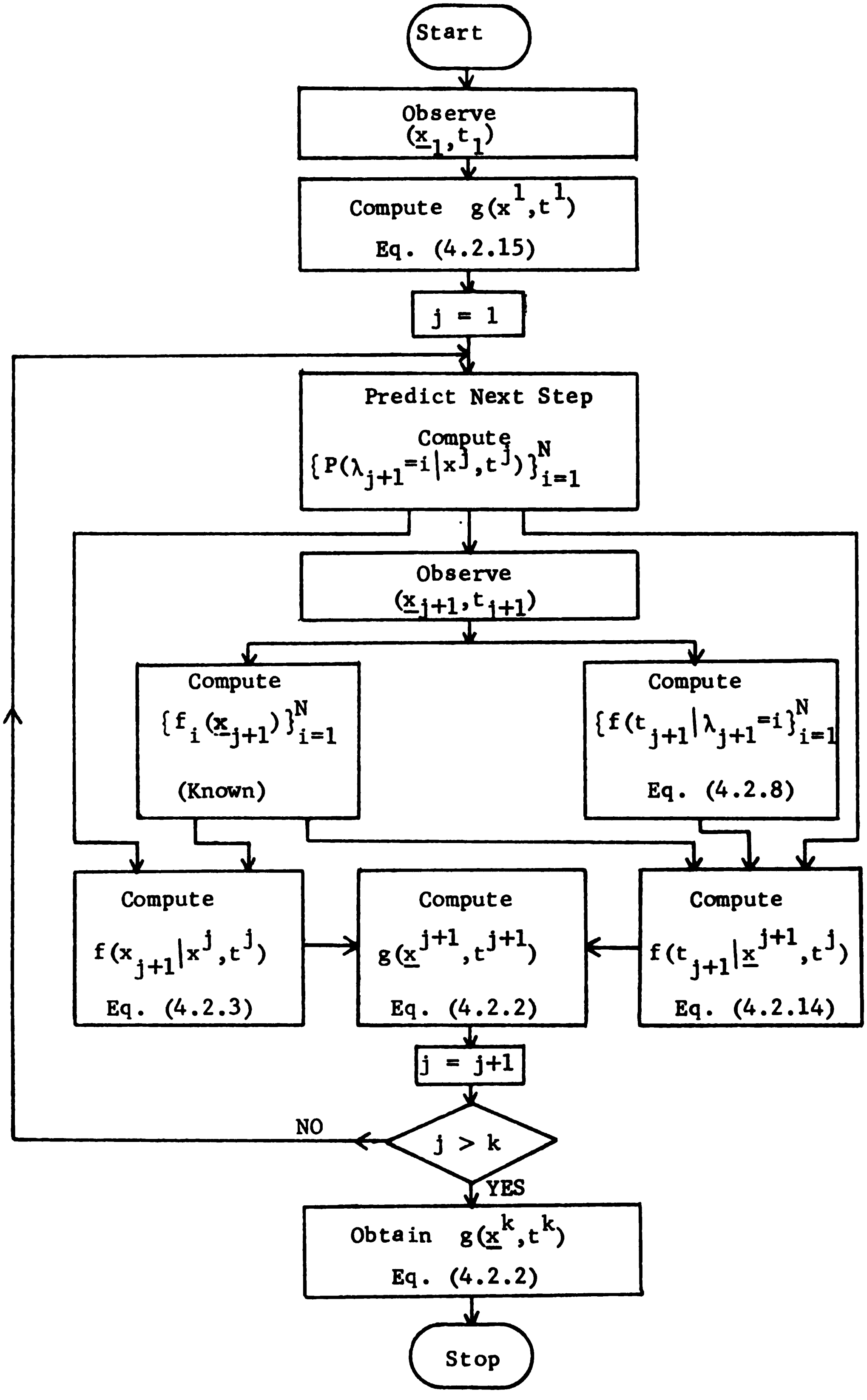

Start
Observe $(\underline{x}_1, t_1)$
Compute $g(x^1, t^1)$
Eq. (4.2.15)
$j = 1$
Predict Next Step
Compute $\{P(\lambda_{j+1}=i \mid x^j, t^j)\}_{i=1}^N$
Observe $(\underline{x}_{j+1}, t_{j+1})$
Compute $\{f_i(\underline{x}_{j+1})\}_{i=1}^N$
(Known)
Compute $\{f(t_{j+1} \mid \lambda_{j+1}=i\}_{i=1}^N$
Eq. (4.2.8)
Compute $f(x_{j+1} \mid x^j, t^j)$
Eq. (4.2.3)
Compute $g(\underline{x}^{j+1}, t^{j+1})$
Eq. (4.2.2)
Compute $f(t_{j+1} \mid \underline{x}^{j+1}, t^j)$
Eq. (4.2.14)
$j = j+1$
NO
$j > k$
YES
Obtain $g(\underline{x}^k, t^k)$
Eq. (4.2.2)
Stop

Fig. 4.2.1: An Algorithm for Iterative Implementation of $g(\underline{x}^k, t^k)$.

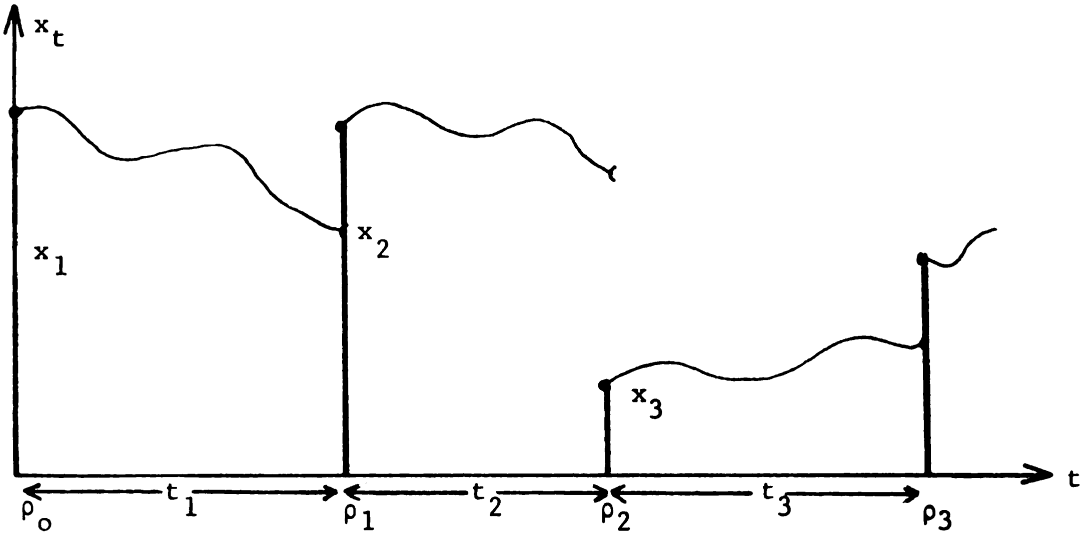


Fig. 4.3.2: A Sampling Scheme

2. Choose processes for which the joint distribution of $\{\underline{x}_k\}_{k=1}^{\infty}$ can be simply described.

In the following developments, the iterative results obtained for optimal decision rule are applied to the case when the random processes $\{x_i(t)\}_{i=1}^{N}$ defining the states of the continuous-parameter Markov chains, are wide-sense stationary normal processes. In this case, the component densities $\{f_i(.)\}_{i=1}^{N}$ for observation vector $\underline{x}_k$, a vector of order n_i, have known forms. The mean vectors $\underline{\mu}_i$ and covariance matrices Σ_i are assumed known. They can be computed from the mean function and auto-correlation function of the corresponding process as follows: $\underline{\mu}_i = \mu_i I$, where I is a unit vector of order n_i and $\mathrm{Cov}(x_{ku}, x_{kv}) = R_i(t_{ku} - t_{kv})$, $u,v = 1,2,\dots,n_i$. The joint density $f_i(.)$ may be written in the following form and is the same for all pattern classes.

$$f_i(\underline{x}_k) = \frac{1}{(2\pi)^{n_i/2}\, |\Sigma_i|^{1/2}} \exp\{-\frac{1}{2}(\underline{x}_k - \mu_k I)^T \Sigma_i^{-1} (\underline{x}_k - \mu_k I)\} \qquad (4.3.1)$$

The iterative, optimal decision rule in Sec. 4.2 can be rewritten as

$$d^*(\underline{x}^k,t^k) = s \quad \text{if} \quad P_s^o g_s(\underline{x}^k,t^k) > P_\ell^o f_\ell(\underline{x}^k,t^k) \quad \ell \neq s \tag{4.3.2}$$

where, from (4.22), the density required has the form

$$g_\ell(\underline{x}^k,t^k) = J_k^\ell(\underline{x}_k,t_{k-1},\{\Gamma_{k-1}^{i\ell}\}_1^N)g_\ell(\underline{x}^{k-1},t^{k-1}) \tag{4.3.3}$$

Omitting, again, supercript and subscript ℓ, J_k^ℓ and $\Gamma_k^{i\ell}$ are defined as:

$$J_k(\underline{x}_k,t_k,\{\Gamma_{k-1}^i\}_1^N) \triangleq J_{k1}(\underline{x}_k,\{\Gamma_{k-1}^i\}_1^N)J_{k2}(\underline{x}_k,t_k,\{\Gamma_{k-1}^i\}_1^N) \tag{4.3.4}$$

$$J_{k1} \triangleq f(\underline{x}_k|\underline{x}^{k-1},t^{k-1}), \quad J_{k2} \triangleq f(t_k|\underline{x}^k,t^{k-1}) \tag{4.3.5}$$

$$\Gamma_k^i = \Gamma_k^i(\underline{x}_k,t_k,\{\Gamma_{k-1}^j\}_1^N) \triangleq (\lambda_{k+1}|\underline{x}^k,t^k) \tag{4.3.6}$$

The following expressions for the Γ_k^i, J_{k1} and J_{k2} are derived using the iterative equations in Sec. 4.2.

$$\Gamma_k^i = \frac{\sum\limits_{\substack{j=1\\ j\neq i}}^N \Gamma_{k-1}^j \frac{r_{ji}\, q_j}{(2\pi)^{n_i/2}|\Sigma_j|^{1/2}} \exp\{-\frac{1}{2}(\underline{x}_k-\mu_j I)^T \Sigma_j^{-1}(\underline{x}_k-\mu_j I) - q_j t_k\}}{\sum\limits_{m=1}^N \Gamma_{k-1}^m \frac{q_m}{(2\pi)^{n_i/2}|\Sigma_m|^{1/2}} \exp\{-\frac{1}{2}(\underline{x}_k-\mu_m I)^T \Sigma_m^{-1}(\underline{x}_k-\mu_m I) - q_m t_k\}}$$

$$J_{k1} = \sum_{i=1}^N \Gamma_{k-1}^i \frac{1}{(2\pi)^{n_i/2}|\Sigma_i|^{1/2}} \exp\{-\frac{1}{2}(\underline{x}_k-\mu_i I)\Sigma_i^{-1}(\underline{x}_k-\mu_i I)\}$$

$$J_{k2} = \frac{\sum\limits_{i=1}^N \Gamma_{k-1}^i \frac{q_i}{(2\pi)^{n_i/2}|\Sigma_i|^{1/2}} \exp\{-\frac{1}{2}(\underline{x}_k-\mu_i I)^T \Sigma_i^{-1}(\underline{x}_k-\mu_i I) - q_i t_k\}}{\sum\limits_{m=1}^N \Gamma_{k-1}^m \frac{1}{(2\pi)^{n_i/2}|\Sigma_m|^{1/2}} \exp\{-\frac{1}{2}(\underline{x}_k-\mu_m I)^T \Sigma_m^{-1}(\underline{x}_k-\mu_m I)}$$

Note that $\{q_i\}_1^N$ and $\{r_{ij}\}_{\substack{i,j=1\\ j\neq i}}^N$ depend on the pattern class, but the mean vectors and covariance matrices are the same for all pattern classes. The initial values of Γ_k^i, J_{k1} and J_{k2} can be easily determined by putting $k = 1$ in the above expressions and noting that $\Gamma_o^i = P_i^o$.

The optimal decision rule can also be established for the case when the random processes $\{x_i(t)\}_{i=1}^N$, describing the states of the chains, are Gauss-Markov processes. Then, the transition probabilities and initial distributions for a given observation vector $\underline{x}_k$ may be written in the form:

$$p_i(x_{k,j+1}|x_{k,j}) = \frac{1}{\sqrt{2\pi(1-R_i)\sigma_i^2}} \exp\left\{- \frac{[x_{k,j+1}-(x_{k,j}-\mu_i)R_i-\mu_i]^2}{2\sigma_i^2(1 - R_i^2)}\right\} \quad (4.3.7)$$

$$P_i^o(x_{k,1}) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left\{- \frac{(x_{k,1} - \mu_i)^2}{2\sigma_i^2}\right\} \quad (4.3.8)$$

Note that the double subscript notation, $x_{k,j}$, was used to denote x_{kj} for convenience. Here, μ_i, σ_i^2 and R_i are the corresponding mean, variance and correlation functions of the process $x_i(t)$ where $R_i = \exp\{-\gamma_i|\tau|\}$; τ is the sampling interval, $\frac{1}{\gamma_i}$ is the correlation time. The joint density of $\underline{x}_k = (x_{k1},\ldots,x_{kn_k})$ is determined in terms of (4.3.7) and (4.3.8) as follows:

$$g_i(\underline{x}_k) = P_i^o(x_{k,1}) \prod_{j=1}^{n_k-1} p_i(x_{k,j+1}|x_{k,j})$$

$$= \frac{P_i^o(x_{k,1})}{A_i} \exp\left\{- \frac{\sum_{j=1}^{n_k-1}[x_{k,j+1} - (x_{k,j}-\mu_i)R_i - \mu_i]^2}{2\sigma_i^2(1 - R_i^2)}\right\}$$

where $A_i \triangleq [2\pi(1 - R_i^2)\sigma_i^2]^{\frac{n_k-1}{2}}$. Then, the necessary quantities, Γ_k^i, J_{1k}, J_{2k} defined in (4.3.4), (4.3.5), (4.3.6), determining the iterative optimal decision rule, are computed. The expressions are given below.

$$\Gamma_k^i =$$

$$\frac{\displaystyle\sum_{\substack{j=1\\ j\neq i}}^{N} \Gamma_{k-1}^{j} \frac{r_{ji} q_j p_j^o(x_{k,1})}{A_j} \exp\left\{-\frac{\sum_{r=1}^{n_k-1}[x_{k,r+1}-(x_{k,r}-\mu_j)R_j - \mu_j]^2}{2\sigma_j^2(1 - R_j)^2} - q_j t_k\right\}}{\displaystyle\sum_{m=1}^{N} \Gamma_{k-1}^{m} \frac{q_m p_m^o(x_{k,1})}{A_m} \exp\left\{-\frac{\sum_{r=1}^{n_k-1} x_{k,r+1}-(x_{k,r}-\mu_m)R_m-\mu_m{}^{2}}{2\sigma_m^2(1 - R_m^2)} - q_m t_k\right\}}$$

$$J_{k1} = \sum_{i=1}^{N} \Gamma_{k-1}^{i} \frac{p_i^o(x_{k,1})}{A_i} \exp\left\{-\frac{\sum_{r=1}^{n_k-1}[x_{k,r+1} - (x_{k,r}-\mu_i)R_i - \mu_i]^2}{2\sigma_i^2(1 - R_i^2)}\right\}$$

$$J_{k2} =$$

$$\frac{\displaystyle\sum_{i=1}^{N} \Gamma_{k-1}^{i} \frac{q_i p_i^o(x_{k,1})}{A_i} \exp\left\{-\frac{\sum_{r=1}^{n_k-1}[x_{k,r+1} - (x_{k,r}-\mu_i)R_i - \mu_i]^2}{2\sigma_i^2(1 - R_i^2)} - q_i t_k\right\}}{\displaystyle\sum_{m=1}^{N} \Gamma_{k-1}^{m} \frac{p_m^o(x_{k,1})}{A_m} \exp - \frac{\sum_{r=1}^{n_k-1}[x_{k,r+1} - (x_{k,r}-\mu_m)R_m - \mu_m]^2}{2\sigma_m^2(1 - R_m^2)}}$$

The initial values of Γ_k^i, J_{k1} and J_{k2} are given as follows:

$$\Gamma_1^i = \frac{\displaystyle\sum_{\substack{j=1\\ j\neq i}}^{N} P_j^o t_{ji} p_j^o(x_{1,1}) q_j \exp\{-q_j t_1\}}{\displaystyle\sum_{m=1}^{N} P_m^o p_m^o\,(x_{1,1}) q_m \exp\{-q_m t_1\}} ;$$

where $A_i \triangleq [2\pi(1 - R_i^2)\sigma_i^2]^{\frac{n_k-1}{2}}$. Then, the necessary quantities, Γ_k^i, J_{1k}, J_{2k} defined in (4.3.4), (4.3.5), (4.3.6), determining the iterative optimal decision rule, are computed. The expressions are given below.

$$\Gamma_k^i = \frac{\displaystyle\sum_{\substack{j=1\\ j\neq i}}^{N} \Gamma_{k-1}^j \frac{r_{ji} q_j p_j^o(x_{k,1})}{A_j} \exp\left\{-\frac{\sum_{r=1}^{n_k-1}[x_{k,r+1}-(x_{k,r}-\mu_j)R_j - \mu_j]^2}{2\sigma_j^2(1 - R_j)^2} - q_j t_k\right\}}{\displaystyle\sum_{m=1}^{N} \Gamma_{k-1}^m \frac{q_m p_m^o(x_{k,1})}{A_m} \exp\left\{-\frac{\sum_{r=1}^{n_k-1} x_{k,r+1}-(x_{k,r}-\mu_m)R_m-\mu_m{}^2}{2\sigma_m^2(1 - R_m^2)} - q_m t_k\right\}}$$

$$J_{k1} = \sum_{i=1}^{N} \Gamma_{k-1}^i \frac{p_i^o(x_{k,1})}{A_i} \exp\left\{-\frac{\sum_{r=1}^{n_k-1}[x_{k,r+1} - (x_{k,r}-\mu_i)R_i - \mu_i]^2}{2\sigma_i^2(1 - R_i^2)}\right\}$$

$$J_{k2} = \frac{\displaystyle\sum_{i=1}^{N} \Gamma_{k-1}^i \frac{q_i p_i^o(x_{k,1})}{A_i} \exp\left\{-\frac{\sum_{r=1}^{n_k-1}[x_{k,r+1} - (x_{k,r}-\mu_i)R_i - \mu_i]^2}{2\sigma_i^2(1 - R_i^2)} - q_i t_k\right\}}{\displaystyle\sum_{m=1}^{N} \Gamma_{k-1}^m \frac{p_m^o(x_{k,1})}{A_m} \exp - \frac{\sum_{r=1}^{n_k-1}[x_{k,r+1} - (x_{k,r}-\mu_m)R_m - \mu_m]^2}{2\sigma_m^2(1 - R_m^2)}}$$

The initial values of Γ_k^i, J_{k1} and J_{k2} are given as follows:

$$\Gamma_1^i = \frac{\displaystyle\sum_{\substack{j=1\\ j\neq i}}^{N} P_j^o t_{ji} p_j^o(x_{1,1}) q_j \exp\{-q_j t_1\}}{\displaystyle\sum_{m=1}^{N} P_m^o p_m^o(x_{1,1}) q_m \exp\{-q_m t_1\}} \; ;$$

$$J_{11} = \sum_{i=1}^{N} P_i^o p_i^o(x_{1,1}) \; ; \; J_{12} = \frac{\sum_{i=1}^{N} P_i^o p_i^o(x_{1,1}) q_i \exp\{-q_i t_1\}}{\sum_{m=1}^{N} P_m^o p_m^o(x_{1,1})}$$

4.4 ADAPTIVE DECISION MAKING

Adaptive decision making schemes are now studied when the underlying Markov model is not completely specified. The unknowns of the model are the Q-matrices, $Q_s = [q_{ij}^{(s)}]_{i,j=1}^{N}$, $q_i^{(s)} \triangleq q_{ii}^{(s)}$, $s \in \Theta$. The parameters and the properties of the random processes, $\{x_i(t)\}_1^N$, describing the states of the underlying continuous-parameter Markov chains are assumed to be known. The component densities, $\{f_i(\cdot|\theta)\}_1^N$, defined in Sec. 4.1, are, then, completely known. Furthermore, for simplicity, they are again chosen to be the same for each pattern class. That is,

$$f_i(\underline{x}_k) = f_i(\underline{x}_k|\theta) \quad \forall \quad \underline{x}_k \qquad (4.4.1)$$

A prior distribution summarizing initial knowledge about the unknown infinitesimal parameters is available for each pattern class. Using a supervised learning scheme, posterior distributions for the fixed, but unknown, parameters are "learned" from training samples $(\underline{y}_s^{n_s}, \tau_s^{n_s})$. The training data are employed to form posterior densities for the parameters which, in turn, lead to an adaptive decision rule which makes decisions with minimum error probability. The purpose of Sec. 4.4.1 is to derive the optimum-adaptive decision rule using the supervised learning scheme. In Sec. 4.2.2, components of the optimum-adaptive decision rule are generated iteratively and

the structure and computational feasibility of the iterative scheme are discussed. The learning feature of the rule is discussed in Sec. 4.4.3.

4.4.1 OPTIMAL-ADAPTIVE DECISION RULE

The optimum-adaptive decision rule will be derived for the problem defined above. The following definitions introduce the problem.

The non randomized decision rule $d \triangleq d(x^k,t^k)$, given the observation sequence (x^k,t^k), given all training samples $(y,\tau) \triangleq \{(\underline{y}_1^{n_1},\tau_1^{n_1}),\dots,(\underline{y}_M^{n_M},\tau_M^{n_M})\}$ and given Q-matrices $\underline{Q} = \{Q_1,Q_2,\dots,Q_M\}$, has expected loss

$$R[d|(\underline{x}^k,t^k),(y,\tau),\underline{Q}] = \sum_{i=1}^{M} L[d(x^k,t^k),i]P[\theta=i|(x^k,t^k),(y,\tau),\underline{Q}] \qquad (4.4.2)$$

Equation (4.4.2) is referred to as the sample-parameter-conditional risk[2]. The parameter-conditional, adaptive-Bayes risk is defined as:

$$r_a(P^o,d|\underline{Q}) = \int_{S_k} R[d|(x^k,t^k),(y,\tau),\underline{Q}]f(\underline{x}^k,t^k|(y,\tau),\underline{Q})d\underline{x}^k\,dt^k \qquad (4.4.3)$$

where $P^o = (P_1^o,P_2^o,\dots,P_M^o)$ is the prior distribution over the parameter space and

$$f[\underline{x}^k,t^k|(y,\tau),Q] = \sum_{i=1}^{M} P_i^o g_i[\underline{x}^k,t^k|(\underline{y}_i^{n_i},\tau_i^{n_i}),\underline{Q}_i] \qquad (4.4.4)$$

Thus, the adaptive-Bayes risk is given by

$$r_a(P^o,d) = \int_{\underline{Q}} r_a(P^o,d|\underline{Q})f_o(\underline{Q})d\underline{Q} \qquad (4.4.5)$$

[2] Signori [S-2], pg. 14.

This development, which involves the use of conditional risks, differs from that given in Appendix B and emphasises the role of prior information in constructing the total risk.

The optimum-adaptive decision rule $d_a^* \triangleq d^*(x^k,t^k)$ is defined as a non-randomized decision rule which minimizes the adaptive-Bayes risk. That is,

$$r_a(P^o,d_a^*) = \inf_{d\in D} r_a(P^o,d) \qquad (4.4.6)$$

where, D is the set of all non-randomized decision rules. Note that

$$r_a(P^o,d_a^*) = E\{R[d_a^*|(\underline{x}^k,t^k),(y,\tau),Q]\} \qquad (4.4.7)$$

In particular, assuming the 0-1 loss function, the optimum-adaptive decision rule becomes a minimum probability of error rule defined by

$$d_a^*(x^k,t^k) = s \quad \text{if} \quad p[\theta=s|(\underline{x}^k,t^k),(\underline{y}_s^{n_s},\tau_s^{n_s})] = \max_{\ell\in\Theta} p[\theta=\ell|(\underline{x}^k,t^k),(y_\ell^{n_\ell},\tau_\ell^{n_\ell})]$$

$$\text{or,} \quad \text{if} \quad P_s^o g_s(x^k,t^k|y_s^{n_s},\tau_s^{n_s}) = \max_{\ell\in\Theta} P_\ell^o g_\ell(x^k,t^k|y_\ell^{n_\ell},\tau_\ell^{n_\ell}) \qquad (4.4.8)$$

The basic elements of the above decision rule are the posterior densities of the observation sequence, given the learning samples. These densities are obtained by averaging the parameter-conditional densities over the posterior densities as in (4.4.9).

$$g_\ell(\underline{x}^k,t^k|\underline{y}_\ell^{n_\ell},\tau_\ell^{n_\ell}) = \int_{Q_\ell} g_\ell(x^k,t^k|\underline{y}_\ell^{n_\ell},\tau_\ell^{n_\ell},Q_\ell)f(Q_\ell|\underline{y}_\ell^{n_\ell},\tau_\ell^{n_\ell})dQ_\ell \quad \ell\in\Theta \qquad (4.4.9)$$

Both factors in the integrand above are generated iteratively in the next section.

4.4.2 ITERATIVE FORM FOR THE DECISION RULE

The first factor in the integrand of (4.4.9) can be implemented iteratively in exactly the same way as in Sec. 4.2, conditioned on the unknown parameter Q_ℓ. The second factor $f(Q_\ell|\underline{y}_\ell^{n_\ell},\tau_\ell^{n_\ell})$ can be implemented iteratively as follows. In the following, the subscript ℓ will be dropped as a convenience; $f(Q|\underline{y}^n,\tau^n)$ is given by the Bayes theorem as follows:

$$f(Q|\underline{y}^n,\tau^n) = \frac{f(\underline{y}^n,\tau^n|Q)f_o(Q)}{f(\underline{y}^n,\tau^n)} \tag{4.4.10}$$

where $f_o(Q)$ is the prior density function over the Q space. Using the Markov dependency between the samples, (4.4.10) can be written as

$$f(Q|\underline{y}^n,\tau^n) = \frac{f(\tau_n|\underline{y}^n,\tau^{n-1},Q)f(\underline{y}_n|\underline{y}^{n-1},\tau^{n-1},Q)}{f(\tau_n|\underline{y}^n,\tau^{n-1})f(\underline{y}_n|\underline{y}^{n-1},\tau^{n-1})}\, f(Q|\underline{y}^{n-1},\tau^{n-1}) \tag{4.4.11}$$

Equation (4.4.11) is the desired result. The last factor on the right side of (4.4.11), $f(Q|\underline{y}^{n-1},\tau^{n-1})$, is the density in the Q space at the $(k-1)\underline{\text{th}}$ stage and is available from the previous step of the iteration scheme. The first and second factors in the same numerator, $f(\tau_n|\underline{y}^n,\tau^{n-1},Q)$ and $f(\underline{y}_n|\underline{y}^{n-1},t^{n-1},Q)$, are the densities directly utilizing the prior knowledge above and can be obtained iteratively as explained in Sec. 4.3. The denominator of (4.4.11) is a normalization constant which assures that $f(Q|\underline{y}^n,\tau^n)$ integrates over the Q space to unity. The parameter-conditional density, $g_\ell(\underline{x}^k,t^k|y_\ell^{n_\ell},\tau_\ell^{n_\ell},Q_\ell)$, is the density function for observation $(\underline{x}^k,t^k)$ corresponding to pattern class ℓ and is given by (4.2.2) (where Q_ℓ was assumed known) as a function of the unknown parameters and the training samples. This term can be generated iteratively in a manner similar to that of Sec. 4.2, but conditioned on knowledge of the unknown parameters. The information about the fixed but unknown parameters $\{Q_\ell\}_1^M$ is summarized by $f(Q_\ell | \underline{Y}_\ell^{n_\ell}, \tau_\ell^{n_\ell})$ which is generated iteratively as explained above and is employed in the decision rule by the averaging procedure given in (4.4.9).

Even though the iterative expressions are available for both posterior densities in the integrand of (4.4.9), it has one major draw-back. The posterior densities for the parameters are not usually reproducing because the class of densities involved are mixtures. In general, two serious problems are immediately encountered when trying to implement this decision rule on a computer. The storage problem refers to the difficulty in allocating computer storage locations for the posterior densities in (4.4.9). The computation time problem refers to the difficulty inherent in performing the number of computations required to change the old posterior densities calculated at $(k-1)^{\underline{th}}$ step, into the new posterior densities at $k^{\underline{th}}$ step. Since the entries of each matrix Q_i are continuous random variables, in general, the amount of storage required to store these densities is infinite.

The only way to make use of the rule is by quantization of the parameter space. Then, each parameter taken to be a discrete random variable so only a finite number of values needed to be stored and updated with each observation. By quantizing fine enough, it is possible to get arbitrarily close to the optimum solution at the expense of increased memory. However, the memory grows exponentially with the dimensionality of the parameter vector $\underset{\sim}{Q}$, making the method feasible only for problems with a small number of states. As a simple example, assume that $M = 2, N = 3$; then, there are $2 \times 9 = 18$ parameters. Using 10 quanta level for each parameter, 10^{18} storage allocations are required for storing posterior densities of the parameters for both classes. One way to get around this problem is to use some sub-optimum methods in which consistent estimators for the parameters are found and, in turn, used in the decision rule as if they were the true values of the parameters.

4.4.3 ASYMPTOTIC OPTIMALITY AND ADAPTATION

In the previous sub-sections, the decision making problem was studied for the case when the Q-matrices describing the pattern classes were unknown. In that case, the optimum-adaptive decision rule $d_a^*(x^k,t^k)$ was used instead of the optimum decision rule $d^*(x^k,t^k)$. To discuss the learning capability of the optimum adaptive decision rule, the following definition is given, [R-2].

DEFINITION 4.4.1 An optimum-adaptive decision rule d_a^* is said to be an asymptotically optimum decision rule d^* relative to $\underset{\sim}{Q}_o = (Q_1^o,Q_2^o,\ldots,Q_M^o)$, the true value of $\underset{\sim}{Q}$, if and only if

$$\lim_{n_1,\ldots,n_M \to \infty} r(p^o,d_a^*) = r(p^o,d^*)$$

That is the Bayes risk in using the optimum-adaptive decision rule, when $\underset{\sim}{Q}$ is unknown, converges as the number of training samples $n_i \to \infty$, $i \in \{1,2,\ldots,M\}$, to the same limit as the Bayes risk when $\underset{\sim}{Q}$ is known and the optimal decision rule is used.

The following theorem, then, insures the asymptotic optimality of the adaptive-optimum decision rule.

THEOREM 4.1.1 d_a^* is asymptotical optimal relative to $\underline{Q}_o$.

A lemma which will be used in the proof of Theorem 4.1.1 is first given.

LEMMA 4.1.1 $$P[\theta=i|(\underline{x}^k,t^k),(\underline{y}^{n_i},\tau^{n_i}),Q_i] \xrightarrow{n_i\to\infty} P[\theta=i|(x^k,t^k),Q_i^o]$$

$$\text{w.p.1} \quad i \in \{1,2,\ldots,M\}$$

The proof of the above lemma is anologous to that given by Signori [S-2] and will be omitted.

PROOF OF THEOREM 4.1.1: For simplicity, the following are defined.

$$u^k \triangleq (\underline{x}^k,t^k) \; ; \; V^{n_i} \triangleq (\underline{y}_i^{n_i},\tau_i^{n_i}) \; , \; V_{n_i} \triangleq (y_{in_i},\tau_{in_i})$$

Then, from (4.4.6) it follows that

$$\begin{aligned}
0 \le r(p^o,d_a^*) - r(p^o,d^*) &= E[R(d_a^*|u^k) - R(d^*|u^k)] \\
&= E[R(d_a^*|u^k - R(d_a^*|u^k, \{V^{n_i}\}_1^M,\underline{Q}] \\
&\quad + E[R(d_a^*|u^k, \{V^{n_i}\}_1^M,\underline{Q}) - R(d^*|u^k,\{V^{n_i}\}_1^M,\underline{Q}] \\
&\quad + E[R(d^*|u^k, \{V^{n_i}\}_1^M,\underline{Q}) - R(d^*|u^k)]
\end{aligned}$$

But,

$$E[R(d_a^*|u^k,\{V^{n_i}\}_1^M,\underline{Q}) - R(d^*|u^k,\{V^{n_i}\}_1^M,\underline{Q}] = r_a(p^o,d_a^*) - r_a(p^o,d^*) \le 0$$

Thus,

$$\begin{aligned}
0 \le r(p^o,d_a^*) - r(p^o,d^*) &\le E[R(d_a^*|u^k) - R(d_a^*|u^k, \{V^{n_i}\}_1^M,\underline{Q}] \\
&\quad + E[R(d^*|u^k, \{V^{n_i}\}_1^M,\underline{Q}) - R(d^*|u^k)]
\end{aligned}$$

Substituting the values of $R(.|.)$ defined in (4.4.2) in the above, it follows that

$$0 \le r(p^o,d_a^*) - r(p^o,d^*)$$

$$\le E\left\{\sum_{i=1}^{M} L(d_a^*,i)P(\theta=i|u^k) - \sum_{i=1}^{M} L(d_a^*,i)P(\theta=i|u^k,v^{n_i},\underset{\sim}{Q})\right\}$$

$$+ E\left\{\sum_{i=1}^{M} L(d^*,i)P(\theta=i|u^k,v^{n_i},\underset{\sim}{Q}) - \sum_{i=1}^{M} L(d^*,i)P(\theta=i|u^k)\right\}$$

$$\le E\left\{\sum_{i=1}^{M} L(d_a^*,i)[P(\theta=i|u^k) - P(\theta=i|u^k,v^{n_i},\underset{\sim}{Q})]\right\}$$

$$+ E\left\{\sum_{i=1}^{M} L(d^*,i)[P(\theta=i|u^k,v^{n_i},\underset{\sim}{Q}) - P(\theta=i|u^k)]\right\}$$

Using Lemma 4.4.1 in the above, the convergence follows. That is,

$$\lim_{n_1,\ldots,n_M\to\infty} [r(d_a^*) - r(d^*)] = 0 \quad \text{w.p.1}$$

As a result, the rule adapts, or converges, to the optimal rule that would be obtained if $\underset{\sim}{Q}_o$ were known.

The conditions under which the posterior distribution of $\underset{\sim}{Q}$, which summarizes all the information about $\underset{\sim}{Q}_o$ contained in $\{y_i^{n_i},\tau_i^{n_i}\}_1^M$, approaches, with probability one, a delta function whose mass is centered about $\underset{\sim}{Q}_o$ were stated in Chapter II. For decision rule (4.4.8), condition 1 follows from (4.1.2). Condition 2 is assumed. Condition 3 is the major requirement for which a strongly consistent estimator for $\underset{\sim}{Q}_o$ (a function of the observations that converges with probability one to $\underset{\sim}{Q}_o$) must be exhibited.

4.5 CONCLUSIONS

Optimal decision making and Bayesian learning with continuous-parameter Markov systems with unobservable states and observable transition times have been the topics of this chapter. The object of the decision rules is to decide which of M continuous-parameter Markov chains is active. The observable quantities were random processes describing the states of the chains and the sojourn times in these states. The general mathematical model, the properties of the observation processes and sampling schemes were defined in Sec. 4.1. Assuming a prior distribution over the pattern classes and that all the parameters of the model were known, the optimal decision rule was defined in (4.2.1) to be that rule in a given class of rules which minimizes the Bayes risk. Its basic components (4.2.2) were generated iteratively. In Sec. 4.3, the analytic results were obtained for the iterative optimal decision rule using special random processes, the normal process and the Gauss-Markov process, to describe the states of the chains.

In Sec. 4.4, the adaptive decision making was studied when the underlying model was not completely specified. The unknowns of the model were the Q-matrices. A prior distribution summarizing the initial knowledge about these unknown parameters was assumed for each pattern class. The fixed but unknown elements of the Q-matrices were learned, using a supervised learning scheme, from the training samples by forming posterior densities for the parameters.

In Sec. 4.4.1, the basic elements of the adaptive decision making problem were defined, such as sample-parameter conditional risk, the parameter conditional adaptive Bayes risk and adaptive-Bayes risk. Assuming the (0-1) loss function, the optimum-adaptive decision rule was defined as a non-randomized decision rule which minimizes the adaptive Bayes risk. Iterative expressions for the optimum-adaptive rule were obtained in Sec. 4.2.2. Some serious problesm encountered in implementing this decision rule on a computer were discussed. It was noted that only a high storage quantization procedure could be used to implement the rule. Asymptotic optimality of the optimum-adaptive decision rule was proved in a manner that exhibits the learning capability of the rule. It was also shown that, under the stated condition, the posterior density of the parameter $\underline{Q}$, which summarizes the knowledge about the true value of parameter $\underline{Q}_o$. converges to a delta function. Thus, $\underline{Q}_o$ is learned and the rule adapts.

CHAPTER V

DECISION MAKING WITH UNOBSERVABLE STATES AND TRANSITION TIMES

In this chapter, the problem of decision making is investigated when the underlying model is completely specified but neither the states nor the transition times can be observed directly. The basic model considered here consists of two pattern classes, each of which is characterized by an N-state, continuous-parameter Markov chain with different stationary transition probability matrices. Every $T < \infty$ seconds, one of the two classes is chosen according to the probability distribution $P^o = (P_1^o, P_2^o)$ and a sample function, λ_t, is generated from the corresponding chain. It is assumed that the features are selected in a medium disturbed by the addition of white Gaussian noise, so that the observed sample function is

$$x_t = \lambda_t + n_t \ , \qquad 0 \leq t \leq T. \tag{5.01}$$

where n_t is a sample function from a white Gaussian process with

$$E(n_t) = 0 \ ; \qquad R(t) = \nu_o \delta(t) \tag{5.02}$$

The decision problem is, as defined in the previous chapter, to determine which pattern class is active based upon observation of x_t. The entire model is illustrated in Fig. 5.01.

In Sec. 5.1, optimal decision making with discrete-time observations is studied and iterative expressions are obtained for

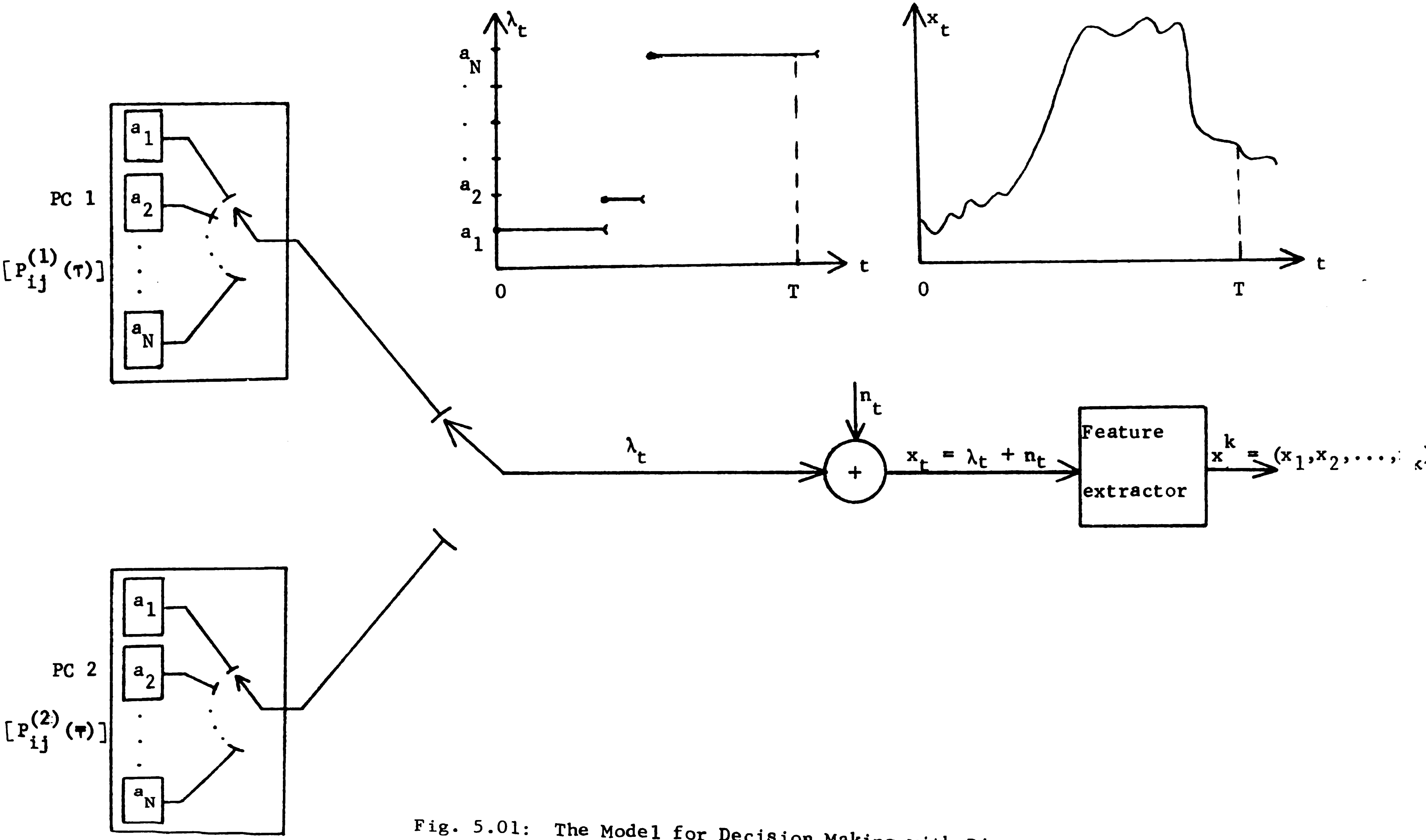


Fig. 5.01: The Model for Decision Making with Discrete Sampling Scheme.

the likelihood ratio. In Sec. 5.2, a continuous likelihood ratio is obtained by a limiting operation. Section 5.3 discusses a method for solving the classical problem of detecting a random telegraph signal in additive white noise.

5.1 OPTIMAL DECISION MAKING WITH TIME SAMPLES

In this section, the features on which decisions are based are defined as point samples of the observed function x_t as follows:

$$x_k = \lambda_k + n_k \ , \quad k = 1,2,\ldots,K, \tag{5.1.1}$$

where $K = T/\tau$ and τ is the interval between samples; $n_k \overset{\Delta}{=} n_{(k-1)\tau}$ represents the noise sample taken from the white Gaussian processes at time $(k-1)\tau$, with

$$E(n_k) = 0, \ \mathrm{Cov}(n_k,n_\ell) = v_0\delta(k-\ell), \ k,\ell = 1,2,\ldots,K. \tag{5.1.2}$$

The sequence of random variables $\{\lambda_k\}_1^K$, $\lambda_k \overset{\Delta}{=} \lambda_{(k-1)\tau}$, take values $\{a_1,a_2,\ldots,a_N\}$ corresponding to a finite state space $\Lambda = \{1,2,\ldots,N\}$. That is, if the Markov process is in state i at time $(k-1)\tau$, then $\lambda_k = a_i$. Since the λ_k's are the time samples from the process λ_t they satisfy the Markov property; namely,

$$P(\lambda_k = a_j|\lambda_{k-1} = a_i,\ldots, \lambda_1 = a_\ell,\theta{=}s) = p_{ij}^{(s)}(\tau) \ \text{ when } \omega_s \text{ is active} \tag{5.1.3}$$

where $p_{ij}^{(s)}(\tau)$, $s = 1,2$, is the stationary transition probability function of the continuous-parameter Markov chain describing chain s.

The observed random variables $\{x_k\}_1^K$ takes values in a finite-dimensional Euclidean space and have the following density when it is known that the chain in state i at time $k\tau$.

$$f_i(x_k) = \rho(x_k - a_i) \tag{5.1.4}$$

Since the noise samples are Gaussian and uncorrelated, it follows that they are state-conditionally independent. That is,

$$\begin{aligned} P(x_{k-1},x_k|\lambda_{k-1} = a_i,\lambda_k = j) &= P(x_{k-1}|\lambda_{k-1} = a_i)P(x_k|\lambda_k = j) \\ &= f_i(x_{k-1})f_j(x_k) \end{aligned} \tag{5.1.5}$$

Since the true states of the chain that is active cannot be observed directly, x_k has the global density:

$$f(x_k|\theta) = \sum_{i=1}^{N} f_i(x_k)P(\lambda_k = a_i|\theta) \qquad \theta \in \Theta \tag{5.1.6}$$

In accordance with the discussion in Chapter III, the optimum decision rule, $d^*(.)$ is given by the Bayes decision rule:

$$\begin{aligned} d^*(x^K) &= \omega_1 \quad \text{if} \quad \Lambda_K < \eta \\ &= \omega_2 \quad \text{if} \quad \Lambda_K > \eta \end{aligned} \tag{5.1.7}$$

where

$$\Lambda_K \overset{\Delta}{=} \Lambda_K(x^K) = \frac{g_2(x^K)}{g_1(x^K)} \; ; \; \eta \overset{\Delta}{=} \frac{P_2^o}{P_1^o} \; ; \; g_s(x^K) = g(x^K|\theta = s)$$

Using conditional probabilities, the likelihood ratio, Λ_K, can be written iteratively as follows:

$$\Lambda_{k+1} = \frac{g_2(x^{k+1})}{g_1(x^{k+1})} = \frac{g_2(x^k)}{g_1(x^k)} \cdot \frac{f(x_{k+1}|x^k,\theta=2)}{f(x_{k+1}|x^k,\theta=1)} \tag{5.1.8}$$

By the total probability law,

$$f(x_{k+1}|x^k,\theta=s) = \sum_{i=1}^{N} f(x_{k+1}|x^k,\lambda_{k+1}=a_i,\theta=s)P(\lambda_{k+1}=a_i|x^k,\theta=s) \quad s = 1,2.$$

By the state-conditional independence, the first factor in the summation above can be written as:

$$f(x_{k+1}|x^k,\lambda_{k+1}=a_i,\theta=s) = f_i(x_{k+1}) = \rho(x_{k+1} - a_i) \qquad (5.1.9)$$

The following notation is now introduced.

$$\Gamma_k^{i1} \triangleq P(\lambda_{k+1}=a_i|x^k,\theta=1) \; ; \; \Gamma_k^{i2} \triangleq P(\lambda_{k+1}=a_i|x^k,\theta=2), \; i \in \Lambda, \; s = 1,2. \qquad (5.1.10)$$

Iterative expressions are then obtained for Λ_K, Γ_k^{i1} and Γ_k^{i2} as follows:

$$\Lambda_{k+1} = \Lambda_K \frac{\sum_{i=1}^{N} \rho(x_{k+1} - a_i)\Gamma_k^{i2}}{\sum_{i=1}^{N} \rho(x_{k+1} - a_i)\Gamma_k^{i1}} \qquad (5.1.11)$$

From the total probability law, Γ_k^{is}, $s = 1,2$, is given by

$$\Gamma_k^{is} = P(\lambda_{k+1}=a_i|x^k,\theta=s) = \sum_{i=1}^{N} P(\lambda_{k+1}=a_i|\lambda_k=a_j,x^k,\theta=s)P(\lambda_k=j|x^k,\theta=s) \qquad (5.1.12)$$

where (5.1.3) and state-conditional independence imply

$$P(\lambda_{k+1}=a_i|\lambda_k=a_j,x^k,\theta=s) = P_{ji}^{(s)}(\tau) \; , \; s = 1,2. \qquad (5.1.13)$$

The second factor in (5.1.12) can be evaluated from the Bayes rule as:

$$P(\lambda_k = a_j | x^k, \theta = s) = \frac{f(x_k | \lambda_k = a_j, \theta = s) P(\lambda_k = a_j | x^{k-1}, \theta = s)}{\sum_{\ell=1}^{N} f(x_k | \lambda_k = a_\ell, \theta = s) P(\lambda_k = \ell | x^{k-1}, \theta = s)}$$

$$= \frac{\rho(x_k - a_j)\Gamma_{k-1}^{js}}{\sum_{\ell=1}^{N} \rho(x_k - a_\ell)\Gamma_{k-1}^{\ell s}} \qquad (5.1.14)$$

Substituting (5.1.13) and (5.1.14) into (5.1.12) gives the following recursive relation.

$$\Gamma_K^{is} = \frac{\sum_{j=1}^{N} p_{ji}^{(s)}(\tau)\rho(x_k - a_j)\Gamma_{k-1}^{js}}{\sum_{\ell=1}^{N} \rho(x_k - a_\ell)\Gamma_{k-1}^{\ell s}}, \quad s = 1,2 \qquad (5.1.15)$$

The initial values, Λ_1 and $\{\Gamma_1^{i1}, \Gamma_1^{i2}\}_1^N$, are given by

$$\Lambda_1 = \frac{\sum_{i=1}^{N} \rho(x_1 - a_i)P_1^o(i)}{\sum_{i=1}^{N} \rho(x_1 - a_i)P_2^o(i)} \qquad (5.1.16)$$

$$\Gamma_1^{is} = \frac{\sum_{i=1}^{N} p_{ji}^{(s)}(\tau)\rho(x_1 - a_j)P_s^o(j)}{\sum_{\ell=1}^{N} \rho(x_1 - a_\ell)P_s^o(\ell)}, \quad s = 1,2, \quad i \in \Lambda. \qquad (5.1.17)$$

where $(P_s^o(1), P_s^o(2), \ldots, P_s^o(N))$; $s = 1,2$, are the prior probabilities over the states.

The recurrent expressions (5.1.11), (5.1.15) and the initial values (5.1.16), (5.1.17) permit sequential computation of the likelihood ratio for any k. The rule $d^*(x^K)$ is an iterative, optimal decision rule which can be though of as an element of a class of on-line decision rules with fixed memory. However, the number of computations needed to make the optimum decision grows linearly with the length of the observation sequence.

5.2 OBSERVATION OF ENTIRE SAMPLE FUNCTION

Expressions analogous to those obtained in Sec. 5.1 can be obtained when the entire sample function x_t is observed for T seconds. In this case, the iterative expressions (5.1.11) and (5.1.15) are replaced by a set of non-linear stochastic differential equations with a limiting argument. Difficulties are encountered when attempts are made to take a mathematically rigorous limit of the results derived from the model of Sec. 5.1 as the sampling interval goes to zero. Because of the infinite variance of the resulting continuous white noise, mathematical operations (e.g., differentiation and integration) do not exist in any strict sense. For this reason, it is necessary to modify the sampling scheme so that the corresponding limits exist. To achieve this, a new observation scheme will be employed.

Let the observed process be $x_t = \lambda_t + n_t$, $t \in [0,T]$, where λ_t is an N-state, continuous-time Markov chain with the stationary transition probability matrix $[p_{ij}^{(s)}(t)]_{i,j=1}^{N}$, $s = 1,2$, and n_t is a white Gaussian process. The process λ_t is first approximated by a homogeneous, N-state, discrete-time Markov process (Markov chain), $\lambda_t^* = \{\lambda_k, k = 1,2,\dots,K\}$, taking values in a finite set $\{a_1,\dots,a_N\}$ at times $(k-1)\tau$; $k = 1,2,\dots,K$, $\tau K = T$, with transition matrix $[p_{ij}^{(s)}(\tau)]_{i,j=1}^{N}$, $s = 1,2$. Good [G-1] showed that λ_t^* will converge to λ_t as $\tau \to 0$ with probability one.

Thus, for sufficiently small τ, x_t may be approximated by

$$x_t \approx \overset{*}{\lambda_t} + n_t \qquad 0 \le t \le T \tag{5.2.1}$$

where the equality will be held when $\tau = 0$, w.p.1. The sampling process being employed here is defined in (5.2.2)

$$x_k = \frac{1}{\tau}\int_{(k-1)\tau}^{k\tau} x_t dt \approx \lambda_k + n_k\, , \quad k = 1,2,\dots,K. \tag{5.2.2}$$

It is clear from (5.02) and (5.2.2) that,

$$E(n_k) = 0, \quad Var(n_k) = \frac{v_o}{\tau} \tag{5.2.3}$$

$$\rho(n_k) = \frac{1}{(2\pi \frac{v_o}{\tau})^{1/2}} \exp\{-\frac{\tau}{2v_o} n_k^2\} \tag{5.2.4}$$

The distributions and the properties of the observed random variables $\{x_k\}_1^K$ are the same as with discrete-time observations and are given in (5.1.4), (5.1.5) and (5.1.6). Thus, the stochastic differential equations for the logarithm of the likelihood ratio, $L_{k+1} = \ell n\Lambda_{k+1}$, can be obtained in the case of continuous-time observations as follows. Taking logarithms in (5.1.11), substituting for $\rho(.)$, and cancelling common terms:

$$L_{k+1} - L_k \approx \ell n \left[\sum_{i=1}^{N} \Gamma_k^{i2} \exp\{-\frac{\tau}{2v_o}(x_{k+1} - a_i)^2\}\right]$$

$$- \ell n \left[\sum_{i=1}^{N} \Gamma_k^{i1} \exp\{-\frac{\tau}{2v_o}(x_{k+1} - a_i)^2\}\right] \tag{5.2.5}$$

For sufficiently small τ,

$$\exp\{-\frac{\tau}{2v_o}(x_{k+1} - a_i)^2\} \approx 1 - \frac{\tau}{2v_o}(x_{k+1} - a_i)^2. \tag{5.2.6}$$

From the definitions of Γ_k^{i1}, Γ_k^{i2}, it follows that

$$\sum_{i=1}^{N} \Gamma_k^{is} = 1 \ , \quad s = 1,2 \ . \tag{5.2.7}$$

Putting these expressions into (5.2.5) produces

$$L_{k+1} - L_k \approx \ell n \left[1 - \frac{\tau}{2v_o} \sum_{i=1}^{N} (x_{k+1} - a_i)^2 \Gamma_k^{i1} \right]$$

$$- \ell n \left[1 - \frac{\tau}{2v_o} \sum_{i=1}^{N} (x_{k+1} - a_i)^2 \Gamma_k^{i1} \right] \ . \tag{5.2.8}$$

Also, for sufficiently small τ,

$$\ell n \left[1 - \frac{\tau}{2v_o} \sum_{i=1}^{N} (x_{k+1} - a_i)^2 \Gamma_k^{is} \right] \approx - \frac{\tau}{2v_o} \sum_{i=1}^{N} (x_{k+1} - a_i)^2 \Gamma_k^{is} \ , \quad s = 1,2. \tag{5.2.9}$$

Thus, (5.2.8) becomes

$$L_{k+1} - L_k \approx \frac{\tau}{2v_o} \sum_{i=1}^{N} (x_{k+1} - a_i)^2 (\Gamma_k^{i1} - \Gamma_k^{i2}). \tag{5.2.10}$$

Dividing both sides by τ and letting $\tau \to 0$, the approximate expression in (5.2.10) becomes an exact stochastic differential equation for the logarithm of the likelihood ratio.

$$\frac{dL_t}{dt} = \frac{1}{2v_o} \sum_{i=1}^{N} (x_t - a_i)^2 (\Gamma_t^{i1} - \Gamma_t^{i2}) \quad 0 \le t \le T \tag{5.2.11}$$

where Γ_t^{is}, $s = 1,2$, is the corresponding value of Γ_k^{is} in continuous time and is defined as

$$\Gamma_t^{is} = P(\lambda_t = a_i | x_\alpha , \ 0 \le \alpha \le t, \ \theta = s) \ , \quad s = 1,2.$$

The initial condition will be obtained from (5.1.16) by setting $\tau = 0$ and taking into account the substitution $L_1 = \ell n \Lambda_1$.

$$L_o \equiv 0 \tag{5.2.12}$$

From (5.2.11) and (5.2.12), it follows that

$$L_t = \frac{1}{2v_o} \sum_{i=1}^{N} \int_o^t (x_z - a_i)^2 (\Gamma_t^{i1} - \Gamma_t^{i2}) dz \qquad (5.2.13)$$

In a similar manner a system of stochastic differential equations are obtained for $\{\Gamma_t^{is}\}_1^N$, $s = 1,2$, as follows. Replacing k by $k+1$ in (5.1.15) and then subtracting Γ_k^{is} from both sides,

$$\Gamma_{k+1}^{is} - \Gamma_k^{is} \approx \frac{\sum_{j=1}^{N} p_{ji}^{(s)}(\tau)\exp\{-\frac{\tau}{2v_o}(x_{k+1} - a_j)^2\}\Gamma_k^{is}}{\sum_{\ell=1}^{N} \exp\{-\frac{\tau}{2v_o}(x_{k+1} - a_\ell)^2\}\Gamma_k^{\ell s}} - \Gamma_k^{is} \; . \quad s = 1,2.$$

For small τ, using approximation (5.2.6) and considering (5.2.7),

$$\Gamma_{k+1}^{is} - \Gamma_k^{is} \approx \frac{\sum_{i=1}^{N} p_{ji}^{(s)}(\tau)\left[1 - \frac{\tau}{2v_o}(x_{k+1} - a_j)^2\right]\Gamma_k^{js}}{1 - \frac{\tau}{2v_o}\sum_{\ell=1}^{N}(x_{k+1} - a_\ell)^2\Gamma^{\ell s}} - \Gamma_k^{is} . \qquad (5.2.14)$$

For sufficiently small τ,

$$\left[1 - \frac{\tau}{2v_o}\sum_{\ell=1}^{N}(x_{k+1} - a_\ell)^2\Gamma_k^{\ell s}\right]^{-1} \approx 1 + \frac{\tau}{2v_o}\sum_{\ell=1}^{N}(x_{k+1} - a_\ell)^2\Gamma_k^{\ell s} . \qquad (5.2.15)$$

Thus, (5.2.14) becomes

$$\Gamma_{k+1}^{is} - \Gamma_k^{is} \approx \left[1 + \frac{\tau}{2v_o}\sum_{\ell=1}^{N}(x_{k+1} - a_\ell)^2\Gamma_k^{\ell s}\right]$$

$$\times \left\{\sum_{j=1}^{N} p_{ji}^{(s)}(\tau)\left[1 - \frac{\tau}{2v_o}(x_{k+1} - a_j)^2\right]\Gamma_k^{js}\right\} - \Gamma_k^{is} .$$

Expanding the right side of the above expression and taking only the first orders terms in τ gives

$$\Gamma_{k+1}^{is} - \Gamma_{k}^{is} \approx -\left(1 - p_{ii}^{(s)}(\tau)\right)\Gamma_{k}^{is} + \sum_{\substack{j=1\\ j\neq i}}^{N} p_{ji}^{(s)}(\tau)\Gamma_{k}^{is}$$

$$- \frac{\tau}{2v_o}\, p_{ii}^{(s)}(\tau)(x_{k+1}-a_i)^2\Gamma_{k}^{is} - \frac{\tau}{2v_o} \sum_{\substack{j=1\\ j\neq i}}^{N} p_{ji}^{(s)}(\tau)(x_{k+1}-a_i)^2\Gamma_{k}^{js}$$

$$+ \frac{\tau}{2v_o}\, p_{ii}^{(s)}(\tau)\Gamma_{k}^{is} \sum_{\ell=1}^{N}(x_{k+1}-a_\ell)^2\Gamma_{k}^{\ell s} + \frac{\tau}{2v_o} \sum_{\substack{j=1\\ j\neq i}}^{N} \sum_{\ell=1}^{N} p_{ji}^{(s)}(\tau)\Gamma_{k}^{\ell s}\Gamma_{k}^{js}(x_{k+1}-a_\ell)^2 .$$

Dividing both sides by τ, letting $\tau \to 0$ and taking into account the following limiting conditions,

$$\lim_{\tau\to 0} \frac{1-p_{ii}^{(s)}(\tau)}{\tau} = q_i^{(s)} \;;\quad \lim_{\tau\to 0} \frac{p_{ij}^{(s)}(\tau)}{\tau} = (1-\Delta_{ij})q_{ij}^{(s)} \;;\quad \lim_{\tau\to 0} p_{ij}^{(s)}(\tau) = \Delta_{ij}, \tag{5.2.16}$$

where Δ_{ij} denotes the Kronecker delta, (5.2.17) follows.

$$\frac{d\Gamma_t^{is}}{dt} = -q_i^{(s)}\Gamma_t^{is} + \sum_{\substack{j=1\\ j\neq i}}^{N} q_{ji}^{(s)}\Gamma_t^{js} - \frac{1}{2v_o}(x_t-a_\ell)^2\Gamma_t^{is} + \frac{1}{2v_o}\Gamma_t^{is}\sum_{\ell=1}^{N}(x_t-a_\ell)^2\Gamma_t^{\ell s}$$

$$i \in \Lambda,\ s = 1,2 . \tag{5.2.17}$$

The initial conditions for $\{\Gamma_t^{is}\}_1^N$, $s = 1,2$, can be obtained from (5.1.17) by putting $\tau = 0$.

$$\Gamma_t^{is} = \frac{P_s^o(i)}{\sum_{j=1}^{N} P_s^o(j)}, \quad i \in \Lambda, \quad s = 1,2. \tag{5.2.18}$$

Expressions (5.2.13), (5.2.17) permit computing the logarithm of the likelihood ratio for any point in time. The stochastic differential equations representing the solutions for L_t and $\{\Gamma_t^{is}\}_1^N$, $s = 1,2$, appear initially to be neat and concise. However, careful examination shows that they represent an infinite-dimensional system of first-order stochastic differential equations, each with the observation process x_t as a driving term. Because of this fact, these stochastic differential equations must be solved by representing the non-linear functions in these equations as series expansions and retaining only the first few terms. The infinite system of differential equations then reduces to a finite set and realizable algorithms are possible [S-6].

5.3 OPTIMAL DETECTION OF RANDOM TELEGRAPH SIGNAL IN ADDITIVE, GAUSSIAN NOISE

The problem of detecting the presence of a random telegraph signal in white Gaussian noise is investigated in this section. This signal is a 2-state, continuous-time Markov chain. Observations are made over a time interval of fixed duration, T. This problem is a special case of the model considered in the previous section. There are two pattern classes. One is associated with the presence of the random telegraph signal and the other pattern class is associated with noise alone.

The optimal decision rule and an iterative implementation are given in Sec. 5.3.1, while the mean and the variance of the logarithm of the likelihood ratio are derived and discussed in Sec. 5.3.2. In Sec. 5.3.3, the probabilities of errors of the first and second kinds are given and recurrent relationships are established for the conditional error probabilities. Finally, in Sec. 5.3.4, stochastic differential equations are obtained both for the continuous logarithm of the likelihood ratio and for the error probabilities.

5.3.1 OPTIMAL DECISION RULE

In the particular statistical decision problem considered here, the action space and the parameter space consist of two elements, $\Theta = \{\theta_o, \theta_1\}$, $\mathcal{A} = \{\omega_o, \omega_1\}$. The basic problem is, then, that of choosing between the two alternatives in (5.3.1); namely,

$$\begin{aligned} \theta_1 &: \; x_k = \lambda_k + n_k \\ \theta_o &: \; x_k = n_k \qquad\qquad k = 1,2,\ldots,K, \end{aligned} \tag{5.3.1}$$

where, $T = \tau K$, τ is the sampling interval, T is the duration of the finite observation interval and K is the (fixed) number of samples. The same sampling scheme is employed as was defined in Sec. 5.1.

The signal process, represented by $\lambda_k = \lambda_{(k-1)\tau}$, $k = 1,2,\ldots,K$, is a 2-state, continuous-time Markov chain which can take values 0 and a with a specified transition-rate matrix Q given by (5.3.2).

$$Q = \begin{bmatrix} -q_o & q_o \\ q_1 & -q_1 \end{bmatrix} \tag{5.3.2}$$

The stationary transition probability matrix, $P(\tau) = [p_{ij}(\tau)]$ can be obtained by solving the backward Kolmogoroff system of differential equations, The complete matrix is:

$$P(\tau) = \begin{bmatrix} \dfrac{q_1}{q_o+q_1} + \dfrac{q_o}{q_o+q_1} e^{-(q_o+q_1)\tau} & \dfrac{q_o}{q_o+q_1} - \dfrac{q_o}{q_o+q_1} e^{-(q_o+q_1)\tau} \\ \dfrac{q_1}{q_o+q_1} - \dfrac{q_o}{q_o+q_1} e^{-(q_o+q_1)\tau} & \dfrac{q_o}{q_o+q_1} + \dfrac{q_1}{q_o+q_1} e^{-(q_o+q_1)\tau} \end{bmatrix} \tag{5.3.3}$$

The noise process, represented by $n_k = n_{(k-1)\tau}$, $k = 1,2,\ldots,K$, is a zero mean, Gaussian noise process.

The observation process $\{x_k\}_1^K$ is as defined in (5.1.1). The optimal decision rule is expressed in terms of the likelihood ratio, Λ_K, and the threshold η. That is,

$$\begin{aligned} &\text{if } \Lambda_K < \eta \quad \text{choose } \omega_o \\ &\text{if } \Lambda_K > \eta \quad \text{choose } \omega_1 \end{aligned} \tag{5.3.4}$$

where the notations are the same as in Sec. 5.1. Taking logarithm of (5.3.5) produces the recurrent relationship for the logarithm of the likelihood ratio,

$$L_k = L_{k+1} + \ln\left[\frac{f(x_k|x^{k-1},\theta_1)}{f(x_k|x^{k-1},\theta_o)}\right] \tag{5.3.5}$$

where $L_k \overset{\Delta}{=} \ln\Lambda_K \quad k = 1,2,\ldots,K.$ (5.3.6)

The probability density functions $f(x_k|x^{k-1},\theta_i)$, $i = 0,1$, are determined from (5.1.6) as

$$f(x_k|x^{k-1},\theta_o) = \rho(x_k) \tag{5.3.7}$$

$$f(x_k|x^{k-1},\theta_1) = \Gamma_{k-1}^o + \Gamma_{k-1}^a\rho(x_k-a) = \rho(x_k)\left[1 + \Gamma_{k-1}^a\left(\frac{\rho(x_k-a)}{\rho(x_k)} - 1\right)\right] \tag{5.3.8}$$

where

$$\Gamma_k^o \overset{\Delta}{=} p(\lambda_{k+1}=0|x^{k+1}) \;;\; \Gamma_k^a \overset{\Delta}{=} p(\lambda_{k+1}=a|x^k) \;;\; \Gamma_k^o + \Gamma_k^a = 1 \;\; \forall \; x^k.$$

The noise density, $\rho(.)$, is defined in (5.1.2). Substituting (5.3.7) and (5.3.8) into (5.3.5),

$$L_k = L_{k+1} + \ln\left[1 + \Gamma_{k-1}^a\left(\frac{\rho(x_k-a)}{\rho(x_k)} - 1\right)\right] . \tag{5.3.9}$$

For the initial value, set $k = 1$.

$$L_1 = \ln\left[1 + p^o(1)\left(\frac{\rho(x_1-a)}{\rho(x_1)} - 1\right)\right] \tag{5.3.10}$$

From arguments similar to those used in the previous section, the following expressions are obtained for Γ_k^o and Γ_k^a, $k = 1,2,\ldots,K$:

$$\Gamma_k^o = \Gamma_k^o(x_k,\Gamma_{k-1}^o,\Gamma_{k-1}^a) = \frac{p_{oo}(\tau)\rho(x_k)\Gamma_{k-1}^o + p_{10}(\tau)R(x_k-a)\Gamma_{k-1}^a}{\rho(x_k)\Gamma_{k-1}^o + \rho(x_k-a)\Gamma_{k-1}^a} \tag{5.3.11}$$

$$\Gamma_k^a = \Gamma_k^a(x_k,\Gamma_{k-1}^o,\Gamma_{k-1}^a) = \frac{p_{o1}(\tau)\rho(x_k)\Gamma_{k-1}^o + p_{11}(\tau)\rho(x_k-a)\Gamma_{k-1}^a}{\rho(x_k)\Gamma_{k-1}^o + \rho(x_k-a)\Gamma_{k-1}^a} \tag{5.3.12}$$

For the initial values, setting $k = 1$ in the above expressions to have

$$\Gamma_1^o = \frac{p_{oo}(\tau)\rho(x_1)p^o(0) + p_{10}(\tau)\rho(x_1-a)p^o(1)}{\rho(x_1)p^o(0) + \rho(x_1-a)p^o(1)} \tag{5.3.13}$$

$$\Gamma_1^a = \frac{p_{o1}(\tau)\rho(x_1)p^o(0) + p_{10}(\tau)\rho(x_1-a)p^o(1)}{\rho(x_1)p^o(1) + \rho(x_1-a)p^o(1)} \tag{5.3.14}$$

5.3.2 ITERATIVE COMPUTATIONS OF THE MEAN AND VARIANCE OF L_k

In this section, the mean and variance of the logarithm of the likelihood ratio are derived. Iterative algorithms are discussed and a theorem, related to the likelihood ratio algorithm, is given.

Before getting into the details, the following is first defined for convenience:

$$\xi_k = \xi_k(x^k) \triangleq \ln\left[1 + \Gamma_{k-1}^a\left(\frac{\rho(x_k-a)}{\rho(x_k)} - 1\right)\right] \tag{5.3.15}$$

The expected value of the lograithm of the likelihood ratio can be determined directly from (5.3.9) as follows:

$$E(L_k - L_{k-1}|\theta_i) = \int_{X^k} \xi_k(x^k)g_i(x^k)dx^k \quad i = 0,1, \tag{5.3.16}$$

In general, it is impossible to evaluate the integral in (5.3.16) analytically. As an alternative, it may help to investigate the conditional expectations, $\mu_\ell(k|\theta_i)$, which are defined as

$$\mu_\ell(k|\theta_i) \overset{\Delta}{=} E(\xi_k|\theta_i,x^\ell) = E\left\{\ell n\left[1 + \Gamma^a_{k-1}\left(\frac{\rho(x_k-a)}{\rho(x_k)} - 1\right)\right]\Bigg| \theta_i,x^\ell\right\},$$

$$\ell = 1,2,\ldots,k$$
$$k = 1,2,\ldots,K$$

Using the above definition and the fact that the σ-field induced by x^ℓ is a sub-field of the σ-field induced by $x^{\ell+1}$, $\mu_\ell(k|\theta_i)$ can be written iteratively as follows:

$$\mu_\ell(k|\theta_i) = E\left\{E\left[\xi_k(k)|x^{\ell+1}\right]\Big|x^\ell\right\}$$
$$= \int_{x_{\ell+1}} \mu_{\ell+1}(k|\theta_i)f(x_{\ell+1}|\theta_i,x^\ell)dx_{\ell+1} \tag{5.3.17}$$

This is the desired result. It is clear that at the final point, i.e., $\ell = k$,

$$\mu_k(k|\theta_i) = \xi_k = \ell n\left[1 + \left(\Gamma^a_{k-1}\ \frac{\rho(x_k-a)}{\rho(x_k)} - 1\right)\right], \quad i = 0,1. \tag{5.3.18}$$

With the help of (5.3.16) and (5.3.17), it is possible to obtain $E(L_k - L_{k-1}|\theta_i)$ by the expression

$$E(L_k - L_{k-1}|\theta_i) = \mu_o(k|\theta_i), \quad i = 0,1. \tag{5.3.19}$$

Substituting the values of $f(x_{\ell+1}|\theta_i,x^\ell)$ into (5.3.17) gives

$$\mu_{\ell}(k|\theta_o) = \int_{x_{\ell+1}} \mu_{\ell+1}(k|\theta_o)\rho(x_{\ell+1})dx_{\ell+1} \tag{5.3.20}$$

$$\mu_{\ell}(k|\theta_1) = \int_{x_{\ell+1}} \mu_{\ell+1}(k|\theta_1)\rho(x_{\ell+1})\left[1 + \Gamma_{\ell}^{a}\left(\frac{\rho(x_{\ell+1}-a)}{\rho(x_{\ell+1})} - 1\right)\right] dx_{\ell+1}. \tag{5.3.21}$$

The variance of the logarithm of the likelihood ratio is now found. Taking the variances of both sides of (5.3.9), when $\theta = \theta_i$, $i = 1,2$, produces the expression for the variance of L_k.

$$\mathrm{Var}(L_k|\theta_i) = \mathrm{Var}(L_{k-1}|\theta_i) + \mathrm{Var}(\xi_k|\theta_i) + 2\ \mathrm{Cov}(L_{k-1},\xi_k|\theta_i) \tag{5.3.22}$$

For the initial value,

$$\mathrm{Var}(L_1|\theta_i) = \mathrm{Var}(\xi_1|\theta_i) \tag{5.3.23}$$

Here, $\mathrm{Var}(L_{k-1}|\theta_i)$ is available from the previous step of iteration scheme. The second and third terms in (5.3.22) are evaluated from the usual definitions of variance and covariance, as

$$\mathrm{Var}(\xi_k|\theta_i) = E(\xi_k^2|\theta_i) - \mu_o^2(k|\theta_i) \tag{5.3.24}$$

$$\mathrm{Cov}(L_{k-1},\xi_k|\theta_i) = E(L_{k-1},\xi_k|\theta_i) - E(L_{k-1}|\theta_i)\mu_o(k|\theta_i) \tag{5.3.25}$$

But, from (5.3.9), L_{k-1} can be expressed as

$$L_{k-1} = \sum_{r=1}^{k-1} \xi_r(x^r) \tag{5.3.26}$$

Substituting this into (5.3.25) gives

$$\mathrm{Cov}(L_{k-1},\xi_k|\theta_i) = \sum_{r=1}^{k-1} E(\xi_r,\xi_k|\theta_i) - \mu_o(k|\theta_i)\sum_{r=0}^{k-1}\mu_o(r|\theta_r) \tag{5.3.27}$$

To simplify the computations of $E(\xi^2|\theta_i)$ and $E(\xi_r,\xi_k|\theta_i)$, the conditional expectations, $s_{\ell}(k|\theta_i)$ and $e_{\ell}(r,k|\theta_i)$, defined below are introduced.

$$s_{\ell}(k|\theta_i) \triangleq E[\xi_k^2(x^k)|\theta_i, x^{\ell}]$$

$$e_{\ell}(r,k|\theta_i) \triangleq E[\xi_r(x^r)\xi_k(x^k)|\theta_i, x^{\ell}]$$

These conditional expectations can be expressed iteratively, using the similar argument as in (5.3.17), as follows:

$$s_{\ell}(k|\theta_i) = \int_{x_{\ell+1}} s_{\ell+1}(k|\theta_i) f(x_{\ell+1}|x^{\ell},\theta_i) dx_{\ell+1}$$

$$e_{\ell}(r,k|\theta_i) = \int_{x_{\ell+1}} e_{\ell+1}(r,k|\theta_i) f(x_{\ell+1}|x^{\ell},\theta_i) dx_{\ell+1}, \quad \ell+1 = 1,2,\ldots,k, \quad i = 0,1.$$

The boundary values are obtained putting $\ell = k$.

$$s_k(k|\theta_i) = \xi_k^2(x^k) = \ell n^2\left[1 + \Gamma_{k-1}^a\left(\frac{\rho(x_k-a)}{\rho(x_k)} - 1\right)\right]$$

$$e_k(r,k|\theta_i) = \xi_r(x^r)\xi_k(x^k) = \ell n\left\{\left[1 + \Gamma_{k-1}^a\left(\frac{\rho(x_k-a)}{\rho(x_k)} - 1\right)\right] \times \left[1 + \Gamma_{r-1}^a\left(\frac{\rho(x_r-a)}{\rho(x_r)} - 1\right)\right]\right\}$$

It is obvious that the usual expected values in (5.3.24) and (5.3.27) are related to these conditional expectations by the expressions:

$$E(\xi_k^2|\theta_i) = s_o(k|\theta_i) \; ; \quad E(\xi_r\xi_k|\theta_i) = e_o(r,k|\theta_i) \; , \quad i = 0,1. \qquad (5.3.28)$$

In terms of (5.3.22), (5.3.24), (5.3.27), (5.3.28) and the boundary values, it is possible to compute $\mathrm{Var}(L_k|\theta_i)$, $i = 0,1$, iteratively.

5.3.3 PROBABILITIES OF ERRORS

The quality of the optimum detector is characterized by the probabilities of errors of the first kind, α_o, and of the second kind, β_o. For the problem considered here, they are defined as follows:

$$\alpha_o = P(L_K > \eta^* | \theta_o), \; \beta_o = P(L_K < \eta^* | \theta_1),$$

where $\eta^* \triangleq \ln\eta$.

In general, exact analytical expressions for α_o and β_o are impossible. However, it is possible to obtain them recursively in terms of the conditional error probabilities, $\alpha_k(K)$ and $\beta_k(K)$ which are defined below, similar to those in (3.4.4), (3.4.5).

$$\alpha_k(K) = P(L_K > \eta^* | \theta_o, x^k), \quad \beta_k(K) = P(L_K < \eta^* | \theta_1, x^k), \; k = 1,2,\ldots,K.$$

By the same argument used in the previous section, the following recurrent expressions for the conditional error probabilities can be obtained.

$$\alpha_k(K) = \int_{x_{k+1}} x_{k+1}(K)\rho(x_{k+1})dx_{k+1} \tag{5.3.29}$$

$$\beta_k(K) = \int_{x_{k+1}} \beta_{k+1}(K)\rho(x_{k+1}) \left[1 + \Gamma_k^a \left(\frac{\rho(x_k - a)}{\rho(x_k)} - 1\right)\right] dx_{k+1} \tag{5.3.30}$$

From (5.3.29) and (5.3.30), it follows that these probabilities of errors at the final point, k = K, must satisfy the expressions:

$$\alpha_K(K) = u(L_k - \eta^*) \; ; \quad \beta_K(K) = 1 - u(L_K - \eta^*),$$

where u(.) is the unit step function. In terms of these relations and the boundary values, the usual error probabilities can be computed by

$$\alpha_o = \alpha_o(K) \ ; \ \beta_o = \beta_o(K)$$

5.3.4 OBSERVATION OF ENTIRE SAMPLE FUNCTION

The results obtained in Sec. 5.3.2 can be specialized to the present problem. The following stochastic differential equation is obtained for the logarithm of the likelihood ratio.

$$\frac{dL_t}{dt} = \frac{1}{2v_o}(2x_t - 1)\Gamma_t^a \qquad 0 \le t \le T \tag{5.3.31}$$

The initial condition is obtained from (5.3.10). Setting $\tau = 0$ in the expressions of L_1,

$$L_o = 0$$

In a similar manner, the following system of non-linear stochastic differential equations are obtained for Γ_t^o and Γ_t^a.

$$\frac{d\Gamma_t^o}{dt} = -q_o\Gamma_t^o + q_1\Gamma_t^a + \frac{1}{2v_o}(1 - 2x_t)\Gamma_t^o\Gamma_t^a \tag{5.3.32}$$

$$\frac{d\Gamma_t^a}{dt} = -q_1\Gamma_t^a + q_o\Gamma_t^o - \frac{1}{2v_o}(1 - 2x_t)\Gamma_t^o\Gamma_t^a \tag{5.3.33}$$

Again, the initial conditions are obtained from (5.3.4) and (5.3.13) by setting $\tau = 0$ in the expressions of Γ_1^o and Γ_1^a.

$$\Gamma_o^o = \frac{p^o(0)}{p^o(0) + p^o(1)} \ ; \quad \Gamma_o^a = \frac{p^o(1)}{p^o(0) + p^o(1)} .$$

The non-linear stochastic differential equation obtained for z_t,

$z_t \overset{\Delta}{=} \Gamma_t^a - \Gamma_t^o$ is identical to that derived in [K-6] by Kulman and Stratonovich, using a completely different technique.

Expressions (5.3.31), (5.3.32), (5.3.33) permit computing the logarithm of the likelihood ratio for any time $t \in [0,T]$. The block diagram of the optimum receiver is realized in terms of these expressions in Fig. 5.3.1.

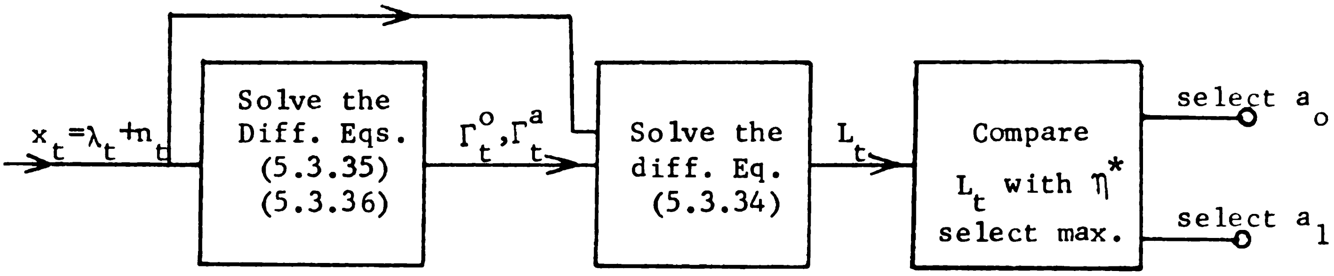

$x_t=\lambda_t+n_t$
Solve the Diff. Eqs. (5.3.35) (5.3.36)
Γ_t^o,Γ_t^a
Solve the diff. Eq. (5.3.34)
L_t
Compare L_t with η^* select max.
select a_o
select a_1

Fig. 5.3.1: The Block Diagram of the Optimum Detector

Recurrent expressions for the conditional probabilities of errors obtained in (5.3.29), (5.3.30) can also be obtained for the case of continuous-time observations. The details of the derivations are as follows: Replacing k with t and k+1 with t+τ, (5.3.29) can be written in a new form.

$$\alpha_t(L_t,\Gamma_t^o) = \int_{x_{t+\tau}} \alpha_{t+\tau}(L_{t+\tau},\Gamma_{t+\tau}^o)\,p(x_{t+\tau})\,dx_{t+\tau} \qquad (5.3.34)$$

Feller [F-1] showed that, under mild regularity conditions, $\alpha \overset{\Delta}{=} \alpha_t(L_t,\Gamma_t^o)$ must be bounded and $p(x_{t+\tau})$ must satisfy the postulates, (4.2), (4.3) and (4.4)[3] and α must satisfy the "backward diffusion equation,"

$$\frac{\partial\alpha}{\partial t} = a_{11}\frac{\partial^2\alpha}{\partial L^2} + 2a_{12}\frac{\partial^2\alpha}{\partial L\partial\Gamma^o} + a_{22}\frac{\partial^2\alpha}{\partial {L^o}^2} + b_1\frac{\partial\alpha}{\partial L} + b_2\frac{\partial\alpha}{\partial\Gamma^o} \qquad (5.3.35)$$

[3] Feller [F-1], pg. 321.

For the problem considered here, it is easily proved that $\rho(x_{t+\tau})$ satisfies the above postulates and since α is a probability of error, then, $|\alpha| \leq 1$. Thus, the regularity conditions are satisfied.

Coefficients in (5.3.35) can be determined as follows: From (5.3.34) for $\tau > 0$, the following equation is first obtained.

$$\frac{\alpha_{t+\tau}(L_t,\Gamma_t^o) - \alpha_t(L_t,\Gamma_t^o)}{\tau} = \frac{\alpha_{t+\tau}(L_t,\Gamma_t^o)}{\tau}$$

$$- \frac{1}{\tau}\int_{x_{t+\tau}} \alpha_{t+\tau}(L_{t+\tau},\Gamma_{t+\tau}^o)\rho(x_{t+\tau})dx_{t+\tau} \qquad (5.3.36)$$

Expanding $\alpha_{t+\tau}(L_{t+\tau},\Gamma_{t+\tau}^o)$ in a Taylor series about (L_t,Γ_t^o) and taking only the first and second order terms gives

$$\alpha_{t+\tau}(L_{t+\tau},\Gamma_{t+\tau}^o) = \alpha_{t+\tau}(L_t,\Gamma_t^o) + L_\tau \frac{\partial\alpha_{t+\tau}(L_t,\Gamma_t^o)}{\partial L_{t+\tau}} + \Gamma_\tau^o \frac{\partial\alpha_{t+\tau}(L_t,\Gamma_t^o)}{\partial\Gamma_{t+\tau}^o}$$

$$+ L_\tau^2 \frac{\partial^2\alpha_{t+\tau}(L_t,\Gamma_t^o)}{\partial L_{t+\tau}^2} + 2L_\tau\Gamma_\tau^o \frac{\partial^2\alpha_{t+\tau}(L_t,\Gamma_t^o)}{\partial L_{t+\tau}\partial z_{t+\tau}} + \Gamma_t^{o2} \frac{\partial^2\alpha_{t+\tau}(L_t,\Gamma_t^o)}{\partial\Gamma_{t+\tau}^{o2}} ,$$

where, $L_\tau \triangleq L_{t+\tau} - L_t$; $\Gamma_\tau^o \triangleq \Gamma_{t+\tau}^o - \Gamma_t^o$.

Substituting this expansion into (5.3.36), performing the resulting integration and then passing to the limit as $\tau \to 0$ produce the coefficients in (5.3.38). The resulting partial stochastic differential equation can be written as:

$$\frac{\partial\alpha}{\partial t} + \frac{1-\Gamma^o}{2v_o}\frac{\partial\alpha}{\partial L} - \left[q_1 + \left(\frac{1}{2v_o} - q_o - q_1\right)\Gamma^o - \frac{1}{2v_o}\Gamma^{o2}\right]\frac{\partial\alpha}{\partial\Gamma^o} - \frac{(1-\Gamma^o)^2}{v_o}\frac{\partial^2\alpha}{\partial L^2}$$

$$+ \frac{(1-\Gamma^o)^2}{v_o}\frac{\partial^2\alpha}{\partial L\partial\Gamma^o} - \Gamma^o(1 - \Gamma^o)\frac{\partial^2\alpha}{\partial\Gamma^{o2}} \qquad (5.3.37)$$

Using similar arguments, a partial stochastic differential equation can be obtained for the conditional probability of error of the second kind and is given below.

$$\frac{\partial\beta}{\partial t} - \frac{(1-\Gamma^o)(1-2\Gamma^o_t)}{2v_o}\frac{\partial\beta}{\partial L} - \left[(\frac{1}{2v_o} - q_o - q_1)\Gamma^o + \frac{1}{2v_o}(1-\Gamma^o)(1-2\Gamma^o)\right]\frac{\partial\beta}{\partial t}$$

$$- \frac{(1-\Gamma^o)^2}{v_o}\frac{\partial^2\alpha}{\partial L^2} + \frac{(1-\Gamma^o)^2}{v_o}\frac{\partial^2\beta}{\partial L\partial\Gamma^o} - \Gamma^{o^2}(1-\Gamma^o)^2\frac{\partial^2\beta}{\partial\Gamma^{o^2}} = 0 \qquad (5.3.38)$$

where

$$\alpha \triangleq \alpha_t(L_t,\Gamma^o_t) \;;\; \beta \triangleq \beta_t(L_t,\Gamma^o_t) \;;\; L \triangleq L_t(x_t,\Gamma^o_t) \;;\; \Gamma^o \triangleq \Gamma^o_t$$

Equations (5.3.37) and (5.3.38) describe the change in a conditional error probability for a recurrent period of time. From (5.3.34), it follows that these probabilities must satisfy the conditions,

$$\alpha_T(L_T,\Gamma^o_T) = u(L_T - \eta^*) \;;\; \beta_T(L_T,\Gamma^o_T) = 1 - u(L_T - \eta^*) \qquad (5.3.39)$$

where T is the time for which the detection is completed. If $\alpha_t(L_t,\Gamma^o_t)$ and $\beta_t(L_t,\Gamma^o_t)$ are the solutions of equations (5.3.37), (5.3.38) with the boundary conditions (5.3.39), then the first kind and the second kind of probabilities of errors are determined by the expressions

$$\alpha_o = \alpha_o(0, \frac{p^o_o}{p^o_o + p^c_o}) \;;\; \beta_o = \beta_o(0, \frac{p^c_o}{p^o_o + p^c_o})$$

5.4 CONCLUSIONS

In the first part of this chapter, the optimal decision making problem was investigated for the model which consists of two pattern classes characterized by different N-state continuous-parameter Markov chains. All parameters of the underlying model were assumed known. Because of the noisy medium which affects the model in an additive manner, neither the states nor the transition times of the sample function generated by these chains could be observed directly.

In Sec. 5.1, assuming additive white Gaussian noise and a discrete-time sampling scheme, the observation process was defined and the Bayes likelihood ratio algorithm was derived. The likelihood ratio and a statistic (5.1.10) were computed recursively. Continuous-time results were obtained by introducing a new sampling scheme and applying a limiting argument. A set of non-linear differential equations were obtained fro the logarithm of the likelihood ratio and for the statistics defined in (5.1.10).

The second part of this chapter considers the problem of detecting a random telegraph signal with a fixed observation time in additive white Gaussian noise. It was shown that the problem is a special case of the model considered in the first part. The optimum detector was defined in Sec. 5.3.1 and its basic components (5.3.9), (5.3.11), were generated recursively. Iterative schemes for computing the mean and variance of the logarithm of the likelihood ratio were given in Sec. 5.3.2. In order to analyze the quality of the optimum detector, it was necessary to consider the probabilities of error of the first and second kind. For this, the conditional error probabilities were first investigated in Sec. 5.3.3. Recurrent relationships were established for them from which the usual error probabilities were found as a function of the observation time and the parameters of the problem.

Finally, in Sec. 5.3.3, the case of continuous observation times were considered. Stochastic differential equations were obtained for the logarithm of the likelihood ratio and for the conditional probabilities of errors.

CHAPTER VI

GENERAL CONCLUSIONS AND EXTENSIONS

The main objectives of the thesis are reviewed in this chapter and possibilities for future research are discussed.

6.1 CONCLUSIONS

This thesis has been concerned with Bayesian decision making and learning algorithms for a particular problem in parametric pattern recognition. Each of M pattern classes was characterized by an N-state, continuous-time, homogeneous Markov chain and the infinitesimal-rate matrices (Q-matrices) for the chains were the parameters of the problem. The object of the decision rules was to decide which of M chains produced the sample function observed. Statistical decision theory was employed throughout the thesis to develop optimal solutions for a special loss function (0-1 loss function).

In Chapter II, the Bayes-optimum decision rules were derived under a perfect observation mechanism (noiseless case). The observable quantities, or the "features", were the sojourn times in the states and the state numbers themselves. Using classified training data from each pattern class, an algorithm for supervised learning was presented and the existence of reproducing prior densities for the parameters was demonstrated. The main result was the formation of recursive, computationally simple parametric forms for the posterior densities of the unknown parameters and for the components of the optimum decision rules. It was shown that the amount of computer storage required to implement these algorithms was fixed and finite. Computer simulations of a specific example gave curves showing probability of error versus the length of the observation sequence for different amounts of training data and demonstrated the inherent practicality of the results.

The problem of computing probability of error was investigated for the noiseless case in Chapter III. All parameters in the model were assumed known and there were only two pattern classes. The exact probability of error, as well as lower and upper bounds, were established for several cases. Conditional error probabilities of the first and second kinds were introduced by which the usual probability of error could be computed iteratively. The main conclusion of the chapter was that, even for a simple case, exact expressions for the probability of error were very complicated and it was difficult to evaluate them. However, the fact that the probability of error decreases toward zero as the number of samples increased without bound was proved by employing a Bhattacharyya bound. The asymptotic behavior of error probability was also studied.

The model of Chapter II was extended in Chapter IV to include the case in which the observations are made in a noisy medium. In the new model, the states of the chains were described by random processes, but sojourn times could be observed. The optimal and the adaptive-optimal decision rules were defined and their basic components were generated iteratively. The asymptotic optimality of the adaptive rules was exhibited. For the known parameter case, the iterative optimal decision rule can be implemented on a computer using only a finite memory. However, computer time increases linearly with the number of observation sequences. One can also conclude that the optimum-adaptive rule is a fixed memory, iterative rule and only a high storage quantization procedure is needed to implement it.

Finally, Chapter V dealt with the case when neither the transition times nor the states could be observed but the processes defining the states were white Gaussian processes. This is equivalent to additive white, Gaussian noise. The Bayes likelihood algorithm was established for the optimal decision rule assuming two pattern classes and a discrete-time sampling scheme. Nonlinear stochastic differential equations were obtained for the logarithm of the likelihood ratio and for the conditional error probabilities when the sampling scheme was changed to permit observation of the entire sample function. The results were applied to the specific problem of detecting a random telegraph signal in white noise. In general, the algorithms for the likelihood ratio are computationally feasible. The implementation of the optimal decision rule with discrete-time sampling requires the knowledge of the stationary transition probability functions for each pattern class. These functions could be obtained by solving N^2 system of linear differential equations with constant coefficients. The continuous sampling scheme leads to much simpler algorithms and do not require transition functions.

6.2 EXTENSIONS

The work in this thesis suggests several extensions that should be investigated.

First is the problem of developing sub-optimum decision rules that are easiér to implement than the high-storage quantization procedure of Sec. 4.4.2 implicit in the optimal-adaptive rule. Strongly consistent estimators for the unknown Q-matrices might be developed. These are functions of the observations that converge with probability one to the true value of the parameter. Conditions for the existence of such estimators are important because, in Sec. 4.4.3, it was shown that a strongly-consistent estimator for $\underset{\sim}{Q}_{o}$ must be exhibited to ensure that these parameters will be learned during the operation of the optimal rule.

Secondly, the expressions obtained for the probability of error can be extended to the cases where M patterns are present and the parameters of the underlying models are not known. The conditional error probabilities, introduced in Sec. 3, suggest a useful computer implementation algorithm for evaluating adaptive decision making devices.

In Chapter V, stochastic differential equations were derived for several quantities, but these equations were not solved. Since, in general, these differential equations are non-linear as well as stochastic, one must not expect to obtain any exact analytical solutions for them. A natural extension of these results would be to seek some approximate solution methods, such as expressing them in the form of difference equations which can be solved on a computer, or, by some reasonable assumptions, reducing them to stochastic linear differential equations for which exact aolutions may exist.

Finally, a particularly significant theorem concerning the Bayes likelihood algorithm in Chapter V can be proved. The probability of error goes to zero as the number of samples, k, increases without bound. This can be proved from the fact that the expected value of the logarithm of the likelihood ratio is a monotonic increasing function of k under θ_1 and a monotonic decreasing function of k under θ_o.

BIBLIOGRAPHY

BIBLIOGRAPHY

[A-1] Anderson, T. W., Introduction to Multivariate Statistical Analysis, Wiley, New York, 1958.

[B-1] Bartlett, M. S., "The Frequency Goodness-of-fit Test for Probability Chains," Proc. Camb. Philos. Soc., Vol. 47, pp. 86-95, 1951.

[B-2] Billingsley, P., "Statistical Methods in Markov Chains," Ann. Math. Stat., Vol. 32, pp. 12-40, 1961.

[B-3] Blackwell, D. and M. A. Girshick, Theory of Games and Statistical Decisions, Wiley, New York, 1954.

[B-4] Braverman, D., "Machine Learning and Automatic Pattern Recognition," Stanford Electronics Laboratories, Technical Report No. 2003-1, Feb., 1961.

[C-1] Chiang, C. L., Introduction to Stochastic Processes in Biostatistics, Wiley, New York, 1968.

[C-2] Chung, L. K., Markov Chains with Stationary Transition Probabilities, Springer-Verlag, Berlin, 1960.

[C-3] Cramér, H., Mathematical Methods of Statistics, Princeton U. Press, 1958.

[D-1] Daly, R. F., "Adaptive Binary Detectors," Stanford Electronic Lab., Technical Report No. 2003-2, June 26, 1961.

[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.

[D-4] Dubes, R. C. and Donoghue, P. J., "Bayesian Learning in Markov Chains with Observable States," Interim Report No. 5, Contract No. AFOSR-1023-67B, Div. Engineering Research, Michigan State University.

[D-5] Dubes, R. C., Hung, A., and McCrum, W. R., "Classification of Electroencephalograms with Pattern Recognition Algorithms," Interim Report No. 4, Contract No. AFOSR-1023-67B, Div. Engineering Research, Michigan State University.

[D-6] Dubes, R. C., The Theory of Applied Probability, Prentice-Hall, Englewood Cliffs, N.J., 1968.

[F-1] Feller, W., An Introduction to Probability Theory and Its Applications, Vol. 2, Wiley, New York, 1966.

[F-2] Ferguson, T. S., Mathematical Statistics: A Decision Theoretic Approach, Academic Press, New York, 1967.

[F-3] Fisher, R. A., Contributions to Mathematical Statistics, Wiley, New York, 1952.

[F-4] Fu, K. S., Sequential Methods in Pattern Recognition and Machine Learning, Academic Press, New York, 1968.

[G-1] Good, I. J., "The Frequency Count of a Markov Chain and the Transition to Continuous Time," Ann. Math, Stat., Vol. , pp. 41-47, October 29, 1960.

[H-1] Hilborn, C. G. and D. G. Lainiotis, "Optimal Estimation in the Presence of Unknown Parameters," IEEE Trans. on System Science and Cybernetic, Vol. Sec-5, January, 1965.

[K-1] Kadota, T. T. and L. A. Shepp, "On the Best Finite Set of Linear Observables for Discriminating Two Gaussian Signals," IEEE Trans. on Information Theory, Vol. IT-13, pp. 278-284, April, 1967.

[K-2] Kailath, T., "A General Likelihood Ratio Formula for Random Signals in Gaussian Noise," IEEE Trans. on Information Theory, Vol. IT-15, pp. 350-361, May, 1969.

[K-3] Kanal, L., "Basic Principle of Some Pattern Recognition Systems," Proc. National Electronic Conference, Vol. 18, pp. 279-295, October, 1962.

[K-4] Karlin, S., A First Course in Stochastic Processes, Academic Press, New York, 1966.

[K-5] Keinosuke, F. and L. G. Koontz, "Application of the Karhunen-Loève Expansion to Feature Selection and Ordering," IEEE Trans. on Computers, Vol. C-19, No. 4, April, 1970.

[K-6] Kulman, N. K., "Certain Optimal Devices for Detection of a Pulse Signal of Random Duration in the Presence of Noise," Radio Engineering and Electronic Physics, Vol. 6, No. 9, pp. 1279-1288, 1961.

[L-1] Loève, M., Probability Theory, VanNostrand, Princeton, N.J., 1963.

[M-1] McLendon, J. R., "A Pseudo Bayes Approach to Digital Detection and Likelihood Ratio Computation," Ph.D. Thesis, Southern Methodist University, 1969.

[M-2] Martin, J. J., Bayesian Decision Problems and Markov Chains, Wiley, New York, 1967.

[N-1] Nagy, G., "State of the Art in Pattern Recognition," Proc. IEEE, Vol. 56, No. 5, May, 1968.

[N-2] Nifontov, Y. A. and V. A. Likharev, "Optimal Detection of a Binary Quantized Markov Signal in the Presence of Noise Similar to the Signal," Engineering Cybernetics, No. 6, 1968.

[N-3] Nillson, N. J., Learning Machines, McGraw-Hill, New York, 1966.

[P-1] Patrick, E. A. and J. C. Hancock, "Learning Probability Spaces for Classification and Recognition of Patterns With or Without Supervision," Purdue University Report TR-EE-65-21, November, 1965.

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

[R-1] Raiffa, H. and R. Schlaifer, Applied Statistical Decision Theory, The M.I.T. Press, Cambridge, Massachusetts, 1961.

[R-2] Raviv, J., "Decision Making in Incompletely Known Stochasic Systems," Int. J. Eng. Sci., Vol. 3, pp. 119-140, Pergamon Press, Long Island City, New York, 1965.

[R-3] Robbins, H., "The Empirical Bayes Approach to Statistical Decision Problems," Ann. Math. Stat., Vol. 35, pp. 1-20, 1964.

[R-4] Royden, H. L., Real Analysis, Macmillan, New York, 1963.

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

[S-2] Signori, D. T., "Estimation and Adaptive Decision Making for Partially Observable Markov Systems," Ph.D. Thesis, Michigan State University, 1968.

[S-3] Silver, E. D., "Markovian Decision Processes with Uncertain Transition Probabilities or Rewards," Ph.D. Thesis, Massachusetts Institute of Technology, 1963.

[S-4] Sosolin, Y. G., "Optimal Detection of Markov Signals and Markov Noise with Discrete Time," Engineering Cybernetics, No. 6, 1966.

[S-5] Spragins, J., "A Note on the Iterative Application of Bayes Rule," IEEE Trans. on Information Theory, Vol. IT-11, No. 4, October, 1965.

[S-6] Stratonovich, R. L., Conditional Markov Processes, American Elsevier Publishing Company, Inc., New York 1968.

[T-1] Takacs, L., Stochastic Processes, Problems and Solutions, Methuen's Monographs on Applied Prob. and Statistics, 1960.

[T-2] Tou, J. T., Computer and Information Sciences-II, Academic Press, New York, 1967.

[T-3] Tucker, H. G., A Graduate Course in Probability Theory, Academic Press, 1967.

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

[W-2] Wassily, H., "Probability Inequalities for Sums of Bounded Random Variables," American Stat. Association J., March, 1963.

[W-3] Watanabe, S., Methodologies of Pattern Recognition, Academic Press, New York, 1969.

APPENDICES

APPENDIX A

DEFINITIONS AND PROPERTIES

In this appendix, some important definitions and theorems related to continuous-parameter Markov chains are given. Most results are based on Doob [D-2] and Chung [C-2].

Consider a family of real random variables indexed by t, $\{x_t, 0 \le t < \infty\}$ on a probability triplet $(\Omega, \mathcal{F}, P)$, where the possible values of x_t is a set of non-negative integers. Here Ω is a probability space, $\mathcal{F}$ is a σ-field of subsets of Ω and P is a probability measure defined on $\mathcal{F}$.

DEFINITION 1.A The stochastic process, $\{x_t, 0 \le t < \infty\}$ is said to be a homogeneous, continuous-time, discrete-state Markov process (continuous- parameter Markov chain), if and only if the following two conditions are satisfied

i) Markov property:

$$P(x_{t_n}=i_n | x_{t_{n-1}}=i_{n-1}, \ldots, x_1=i_1) = P(x_{t_n}=i_n | x_{t_{n-1}}=i_{n-1}) \qquad (A.1)$$

for every $n \ge 2$, $0 \le t_1 < t_2 < \ldots < t_n$ and for all possible values of the random variables in question.

ii) Homogeneity property:

The chain is said to be Homogeneous, or said to have stationary transition probabilities, if and only if

$$P(x_{t+h}=j | x_h=i) = P_{ij}(t) \quad \forall\, i,j \in \Lambda \text{ and } t > 0 \qquad (A.2)$$

where $P_{ij}(t)$ is independent of $h \geq 0$.

DEFINITION 2.A The functions $\{P_{ij}(t)\}_{i,j=1}^{N}$ are called stationary transition probability functions if and only if the following conditions are satisfied:

A1-1. $P_{ij}(t) \geq 0$

A1-2. $\sum_{j=1}^{N} P_{ij}(t) = 1 \quad t > 0, \quad \forall\, i \in \Lambda$

A1-3. $\sum_{k=1}^{N} P_{ik}(t)P_{kj}(s) = P_{ij}(t + s) \quad s,t > 0$

The following continuity condition is also assumed to be satisfied for $t > 0$

A1-4. $\lim_{t\to 0} P_{ij}(t) = \Delta_{ij}$

The matrix $[P_{ij}(t)]_{N\times N}$ is called the stationary transition probability matrix.

DEFINITION 3.A The process $\{x_t;\ 0 \leq t < \infty\}$, is said to be separable (relative to the class of closed sets), if and only if there exists a denumerable subset, $R \subseteq \{t;\ 0 \leq t < \infty\}$, called a separability set, and a null set N with the following property: for any closed linear set A and open interval G in $(-\infty,\infty)$

$$\{\omega\colon X_t(\omega) \in A,\ t \in G \cap R\} - \{\omega\colon X_t(\omega) \in A,\ t \in G \cap [0,\infty)\} \subset N$$

The main results in this thesis are based upon the following theorems which are given without proof:

THEOREM 1.A If $[P_{ij}(t)]$ is the stationary transition probability matrix, then, the following limits exist:

$$\lim_{t\to 0} \frac{1 - P_{ii}(t)}{t} = -P'_{ii}(0) < \infty \quad \forall\, i \in \Lambda,\ q_i \overset{\Delta}{=} P'_{ii}(0) \qquad \text{(A.3)}$$

$$\lim_{t\to 0} \frac{P_{ij}(t)}{t} = P'_{ij}(0) < \infty \quad \forall\, i,j \in \Lambda \ (j \neq i); \ q_{ij} \triangleq P'_{ij}(0) \qquad \text{(A.4)}$$

Furthermore, if $\{x_t : 0 \le t < \infty\}$ is a separable process determined by $[P_{ij}(t)]_{i,j=1}^{N}$ together with an initial probability distribution, $\{P_i ; i \in \Lambda\}$ over the states, then, the probability of remaining in state i for α time units, given that the chain in state i is:

$$P\{x_t = i, h \le t \le h + \alpha | x_h = i\} = e^{-q_i \alpha}, \quad \alpha \ge 0 \qquad \text{(A.5)}$$

In addition, if $q_i > 0 \ \forall\, i \in \Lambda$ and if $x_{t_o} = i$, there is a sample function discontinuity for some $t > t_o$ with probability one (w.p.1). If $0 < \alpha \le \infty$, and if there is a discontinuity in the interval $[t_o, t_o + \alpha]$, the probability that the first jump is to j is q_{ij}/q_i.

The matrix $Q = [q_{ij}]_{i,j=1}^{N}$ where $q_{ii} \triangleq -q_i$, defined above is called the infinitesimal matrix or the transition-rate matrix or the Q-matrix. From A1-1 and A1-2 it follows that

$$q_i \ge 0 \ , \ \sum_{\substack{j=1 \\ j\neq i}}^{N} q_{ij} = q_i \quad \forall\, i,j \in \Lambda \ \ (j \neq i) \qquad \text{(A.6)}$$

Differentiating A1-3 with respect to each variable and setting the result equal to zero produces the following system of differential equations for the stationary transition probability functions which is called the "backward differential equation system".

$$P'_{ik}(t) = -q_i P_{ik}(t) + \sum_{\substack{j=1 \\ j\neq i}}^{N} q_{ij} P_{jk}(t) \quad \forall\, i,k \in \Lambda \qquad \text{(A.7)}$$

The initial conditions are given by

$$P_{ij}(0) = \Delta_{ij} \qquad \forall\ i,j \in \Lambda \tag{A.2}$$

The following question is of interest: What are the necessary and sufficient conditions that the q_i's and q_{ij}'s must satisfy to determine the $\{P_{ij}(t)\}_{i,j=1}^{N}$ uniquely? Or, in other words, what conditions on the q_i's and q_{ij}'s ensure a unique solution to the backward differential equation system? Doob [D-2] shows that the necessary and sufficient conditions under which an allowable set of stationary transition probabilities are defined by the backward differential equation system for a given infinitesimal Q-matrix are that the process is separable (which assures that the sojourn times in the states are independent and exponentially distributed), has a finite number of states, satisfies (A.6), and that the states are stable (i.e., $q_i > 0$). For a discrete-parameter Markov chain, these conditions imply that the process is ergodic.

As long as the above conditions are satisfied, a unique continuous-parameter Markov chain with the stationary transition probability matrix $[P_{ij}(t)]$ can always be found from a given Q-matrix. The process, constructed in this way, is called a minimal process. The relation between the $P_{ij}(t)$'s and q_{ij}'s is given by one of the following integral equations.

$$P_{ik}(t) = \delta_{ik}\, e^{-q_i t} + \sum_{\substack{j=1\\ j\neq i}}^{N} \int_0^t e^{-q_i s}\, q_{jk}\, P_{jk}(t-s)ds \tag{A.9}$$

or

$$P_{ik}(t) = \delta_{ik}\, e^{-q_k t} + \sum_{\substack{j=1\\ j\neq k}}^{N} \int_0^t P_{ij}(s)\, q_{jk}\, e^{-q_k(t-s)}\, ds. \qquad (A.10)$$

If the eigenvalues of the Q-matrix are real and distinct, then the solutions of the above integral equations are the same and are given by

$$P_{ik}(t) = \sum_{\ell=1}^{N} \frac{A^T_{ik}(\ell)\, e^{\rho_\ell t}}{\sum\limits_{\substack{m=1\\ m\neq\ell}}^{N} (\rho_\ell - \rho_m)} \quad \forall\ i,j \in \Lambda, \qquad (A.11)$$

where $\rho_1, \rho_2, \ldots, \rho_N$ are the eigenvalues of the Q-matrix and $A_{ik}(\ell)$ is the cofactor of the matrix $A(\ell) = [\rho_\ell I - Q]$. Here, I is the unit matrix.

Doob also proves that the sample functions of a continuous-parameter Markov chain satisfying the conditions stated above are almost all step functions. The corresponding continuous-parameter Markov chains can thus be constructed in terms of their sample functions in the following manner:

Let $x_1, x_2, \ldots$ be random variables taking values in $\Lambda = \{1,2,\ldots,N\}$ only. Let $T_1, T_2, \ldots$ be positive random variables and suppose $\{q_i\}_{i=1}^N$, $\{q_{ij}\}_{\substack{i,j=1\\ j\neq i}}^N$ are any real positive numbers satisfying (A.6) and $0 < q_i < \infty$. Define the following conditional probabilities:

$$P(T_1 \le t_1 | x_1 = i) = 1 - e^{-q_i t_1} \qquad t_1 > 0 \qquad (A.12)$$

$$P(x_k = j | x_{k-1} = i, \ldots, x_1 = \ell, t_{k-1}, \ldots, t_1) = \frac{q_{ij}}{q_i} \quad q_i > 0\ \forall\ i,j \in \Lambda \quad (j \neq i) \qquad (A.13)$$

$$P(T_k \le t_k | t_{k-1},\dots,t_1,x_k,\dots,x_1) = 1 - e^{-q_j t_k} \qquad t_k > 0 \qquad k \in \Lambda \qquad (A.14)$$

This shows that the random variables $\{x_k\}_{k=1}^{\infty}$ have a Markov dependency and the random variables $\{T_k\}_{k=1}^{\infty}$ are conditionally independent. Now, define the process, $\{x_t : 0 \le t < \infty\}$ as follows:

$$x_t = x_n \qquad \text{if} \qquad \rho_{n-1} \le t < \rho_n$$

where $T_n \overset{\Delta}{=} \rho_n - \rho_{n-1}$ and $\rho_o \equiv 0$. Then, it is easy to show that the process defined in this way is a continuous-parameter Markov chain and that the q_i's and q_{ij}'s satisfy the limit conditions specified in Theorem A.1. The x_n's are the state numbers and T_n's are the sojourn times in these states. A typical sample function from this process is illustrated in Fig. A.1.

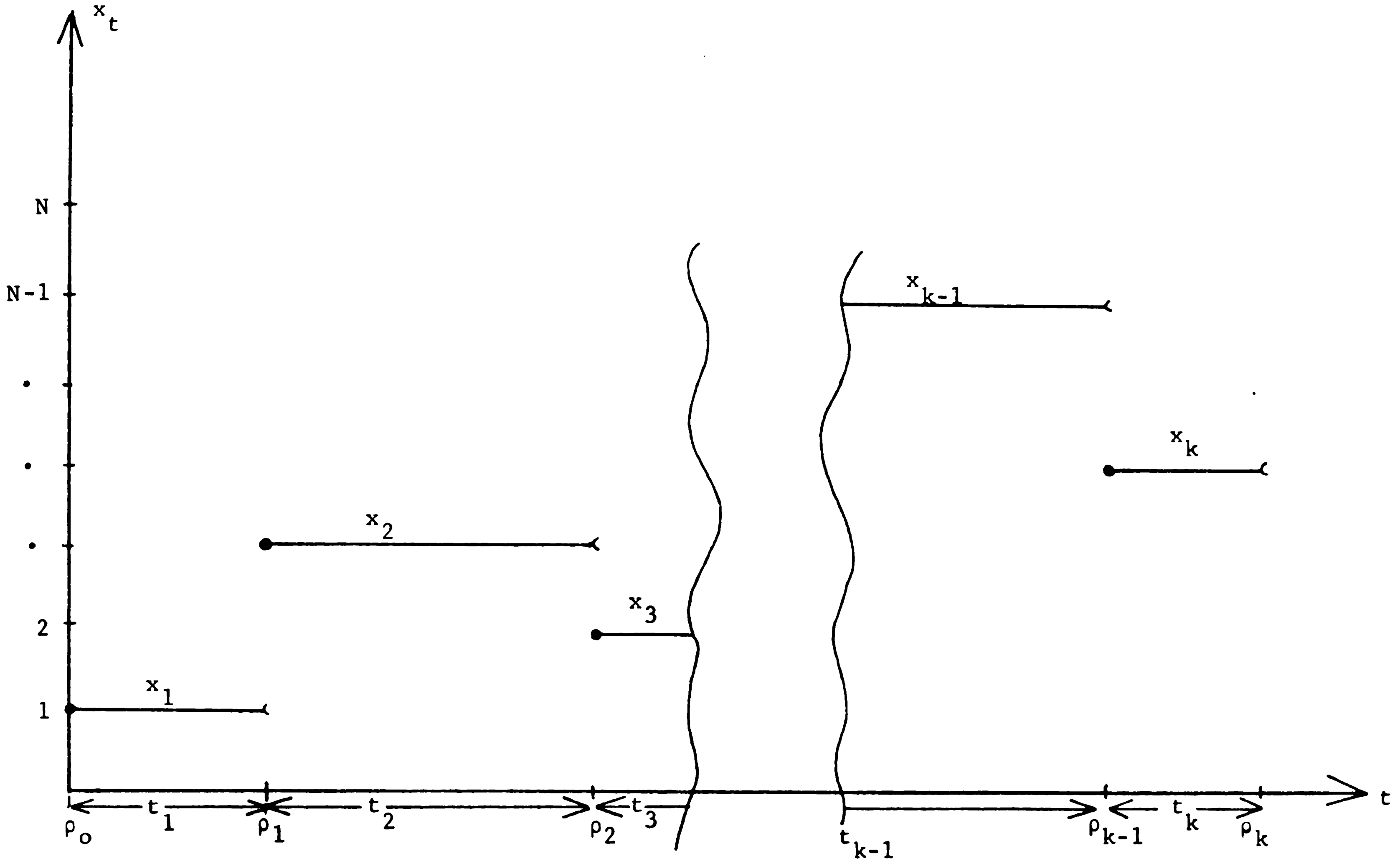


Fig. A.1: Typical sample function from a continuous-time Markov Chain.

APPENDIX B

DERIVATION OF OPTIMAL DECISION RULES

The basic elements of statistical decision theory are summarized in this Appendix. Most of the results are taken from the literature, but are presented from a slightly different point of view, more suitable for the problem of interest in this thesis.

A problem in decision theory consists of three basic elements:

1. A non-empty set, Θ, of possible state of nature (parameter space)
2. A non-empty set, $\mathcal{A}$, called an action space
3. A real-valued loss function, $L(\theta,a)$, defined on $\Theta \times \mathcal{A}$.

In pattern recognition terminology, a state of nature is called a "pattern class". Before making a decision, a random vector, $(x^k,t^k) \triangleq (x_1,\ldots,x_k,t_1,\ldots,t_k)$ is observed whose distribution depends on the true state of nature $\theta \in \Theta$. The sample space, $S_k \triangleq \{\omega: \omega = \left(\prod_{i=1}^{k} \xi_i\right) \times \left(\prod_{i=1}^{k} \eta_i\right)\}$, is taken to be a Borel subset of a finite dimensional Euclidean space and the probability distributions of (x^k,t^k) are defined on the class of Borel subsets $\mathcal{B}$ on S_k. For each $\theta \in \Theta$, there is a product measure $\mu(x^k) \times v(t^k)$ defined on $\mathcal{B}$ and corresponding multivariate density function $g(x^k,t^k|\theta)$, which represents the distribution of (x^k,t^k) when θ is the true state of nature.

On the basis of the set of observations (x^k,t^k), an action $d(x^k,t^k) \in \mathcal{A}$ is chosen. In choosing $d(x^k,t^k)$, the loss $L[\theta,d(x^k,t^k)]$, $\theta \in \Theta$, which is a random quantity, is incurred. The expected value of $L[\theta,d(x^k,t^k)]$, when θ is the true state of nature is called the risk.

$$R(\theta,d) = E_\theta L[\theta,d(x^k,t^k)] \tag{B.1}$$

Any function, $d(.)$, that maps the sample space, S_k, into $\mathcal{A}$ is called a non-randomized decision rule, provided the risk function, $R(\theta,d)$, exists and is finite. The Bayes principle involves the notion of the Bayes risk. A distribution P^o is imposed on the parameter space called a prior distribution. The Bayes risk is the expectation of the risk function with respect to P^o; namely,

$$r(P^o,d) = ER(Z,d) \tag{B.2}$$

where Z is a random variable over Θ having distribution P^o.

DEFINITION B.1 A non-randomized decision rule, $d^*(x^k,t^k)$ is said to be Bayes with respect to the prior distribution P^o, if and only if

$$r(P^o,d^*) = \inf_d r(P^o,d) \tag{B.3}$$

DEFINITION B.2 A non-randomized decision rule $d^*(x^k,t^k)$ is said to be an optimum decision rule if and only if it is Bayes with respect to the prior distribution P^o.

In the following, the optimum decision rule is first derived for the general loss function, and then for the special "0-1" loss function, which leads to a minimum probability of error rule.

In Chapter II, the parameter space, Θ and the action space, $\mathcal{A}$ are finite, i.e., $\Theta = \mathcal{A} = \{1,2,\ldots,M\}$, where each integer in this set corresponds to a different continuous-parameter Markov chain. The random vector (x^k,t^k) is observed. Its value depends on which continuous-parameter Markov chain is active. The problem, then, for any given prior distribution P^o on Θ, is to find a non-randomized decision rule $d^*(x^k,t^k)$ that minimizes the Bayes risk

$$r(P^o,d) = \sum_{i=1}^{M} R(\theta = i,d)P_i^o$$

where

$$R(\theta,d) = \int_{\Lambda^k \times R^k} L[\theta,d(x^k,t^k)]g(x^k,t^k|\theta)d(\mu(x^k) \times \nu(t^k)) \qquad \theta \in \Theta$$

Using Fubini's theorem,

$$R(\theta,d) = \int_{\Lambda^k}\left[\int_{R^k} L[\theta,d(x^k,t^k)]g(x^k,t^k|\theta)d\mu(t^k)\right]d\nu(x^k)$$

Since x^k is defined on $\Lambda^k = \{1,2,\ldots,N\}^k$, the outer Lebesgue integral with respect to counting measure $\nu(x^k)$ becomes a summation over Λ^k, and the inner integral becomes a Riemann integral over $[0,\infty)^k$. Thus,

$$R(\theta,d) = \sum_{\Lambda^k}\int_{[0,\infty)^k} L[\theta,d(x^k,t^k)]g(x^k,t^k|\theta)dt^k \quad \text{where} \quad dt^k \triangleq dt_1dt_2\ldots dt_k$$

Then, the Bayes risk is,

$$r(P^o,d) = \sum_{i=1}^{M}\left[\sum_{\Lambda^k}\int_{[0,\infty)^k} L[\theta=i,d(x^k,t^k)]g(x^k,t^k|\theta=i)dt^k\right]P_i^o$$

Since P^o on Θ is finite, the order of integration and summation can be interchanged. Since $P_i^o\, g(x^k,t^k|\theta=i) = P(\theta=i|x^k,t^k)g(x^k,t^k)$

$$r(P^o,d) = \sum_{\Lambda^k} \int_{[0,\infty)^k} \left[\sum_{i=1}^{M} L(\theta=i,d(x^k,t^k)]P(\theta=i|x^k,t^k)\right] g(x^k,t^k)dt^k \quad \text{(B.4)}$$

Now, to find a function $d^*(.)$ that minimizes $r(P^o,d)$, an action, call it $d^*(x^k,t^k)$, may be found for each $(x^k,t^k) \in S_k$ that minimizes

$$\sum_{i=1}^{M} L[\theta = i,d(x^k,t^k)]P(\theta = i|x^k,t^k).$$

In other words, the optimum decision rule, $d^*(x^k,t^k)$, minimizes the posterior conditional expected loss, given (x^k,t^k).

For the problem considered in Chapter IV, the optimum decision rule is determined in exactly the same way as above. The only difference is that x^k is defined on R^k instead of Λ^k so the counting measure becomes a Lebesgue measure $\mu(x^k)$ and the summation becomes a Riemann integral over R^k.

In particular, when the loss function is given by $L(i,j) = 1 - \Delta_{ij}$ (0-1 loss function), the optimal decision rule obtained above becomes a minimum probability of error rule, defined by

$d^*(x^k,t^k) = s$ if s is the first index such that

$P(\theta = s|x^k,t^k) \geq P(\theta = \ell|x^k,t^k) \quad \forall\, \ell \neq s$

or $\quad P_s^o g(x^k,t^k|\theta=s) \geq P_\ell^o g(x^k,t^k|\theta=\ell) \; \forall\, s \neq \ell \quad s,\ell \in \{1,2,\ldots,M\}$

(B.5)

APPENDIX C.

A PROOF OF CONVERGENCE

The fact that $\psi(q_{ij}|y^n,\tau^n) \xrightarrow{n\to\infty} \delta(q_{ij} - q^o_{ij})$ w.p. 1, $\forall\ i,j \in \Lambda\ (j \neq i)$, is proved in this Appendix. It was shown in Chapter II that

$$f(q_i,r_{ij}|y^n,\tau^n) \xrightarrow{n\to\infty} \delta(q_i - q^o_i,q_{ij} - q^o_{ij}) \quad i,j \in \Lambda\ (j \neq i) \quad \text{(C.1)}$$

But, from the definition of the delta dirac function, $\delta(.)$, it follows that

$$\delta(q_i - q^o_i,q_{ij} - q^o_{ij}) = \delta(q_i - q^o_i)\delta(q_{ij} - q^o_{ij}) \ . \qquad \text{(C.2)}$$

Thus, taking limit of $\psi(q_{ij}|y^n,\tau^n)$ in Sec. 2.6 gives

$$\lim_{n\to\infty} \psi(q_{ij}|y^n,\tau^n) = \lim_{n\to\infty} \int_o^1 \frac{1}{r_{ij}} f(r_{ij}, \frac{q_{ij}}{r_{ij}}|y^n,\tau^n)dr_{ij} \qquad \text{(C.3)}$$

Since $\psi(q_{ij}|y^n,\tau^n) < \infty\ q_{ij},\ 0 < q_{ij} < 1$ and $\forall(y^n,\tau^n) \in S_n$, the limit and integration sign in (C.3) can be interchanged. Doing so,

$$\lim_{n\to\infty} \psi(q_{ij}|y^n,\tau^n) = \int_o^1 \frac{1}{r_{ij}} \lim_{n\to\infty} f(r_{ij}, \frac{q_{ij}}{r_{ij}}|y^n,\tau^n)dr_{ij}. \qquad \text{(C.4)}$$

In terms of (C.1) and (C.2), the last factor in the integrand of (C.4) can be written as:

$$\lim_{n\to\infty} f(r_{ij}, \frac{q_{ij}}{r_{ij}}|y^n,\tau^n) = \delta(r_{ij} - r^o_{ij}, \frac{q_{ij}}{r_{ij}} - q^o_i) = \delta(r_{ij} - r^o_{ij})\delta(\frac{q_{ij}}{r_{ij}} - q^o_i).$$

Substituting this into (C.3) to give

$$\lim_{n\to\infty} \psi(q_{ij}|y^n,\tau^n) = \int_o^1 \frac{1}{r_{ij}} \delta(r_{ij} - r^o_{ij})\delta(\frac{q_{ij}}{r_{ij}} - q^o_i)dr_{ij}, \; 0 < r_{ij} < 1.$$

Define $h(r_{ij}) \triangleq \frac{1}{r_{ij}} \delta(\frac{q_{ij}}{r_{ij}} - q^o_i)$ then,

$$\lim_{n\to\infty} \psi(q_{ij}|y^n,\tau^n) = \int_o^1 h(r_{ij})\delta(r_{ij} - r^o_{ij})dr_{ij} = h(r^o_{ij})$$

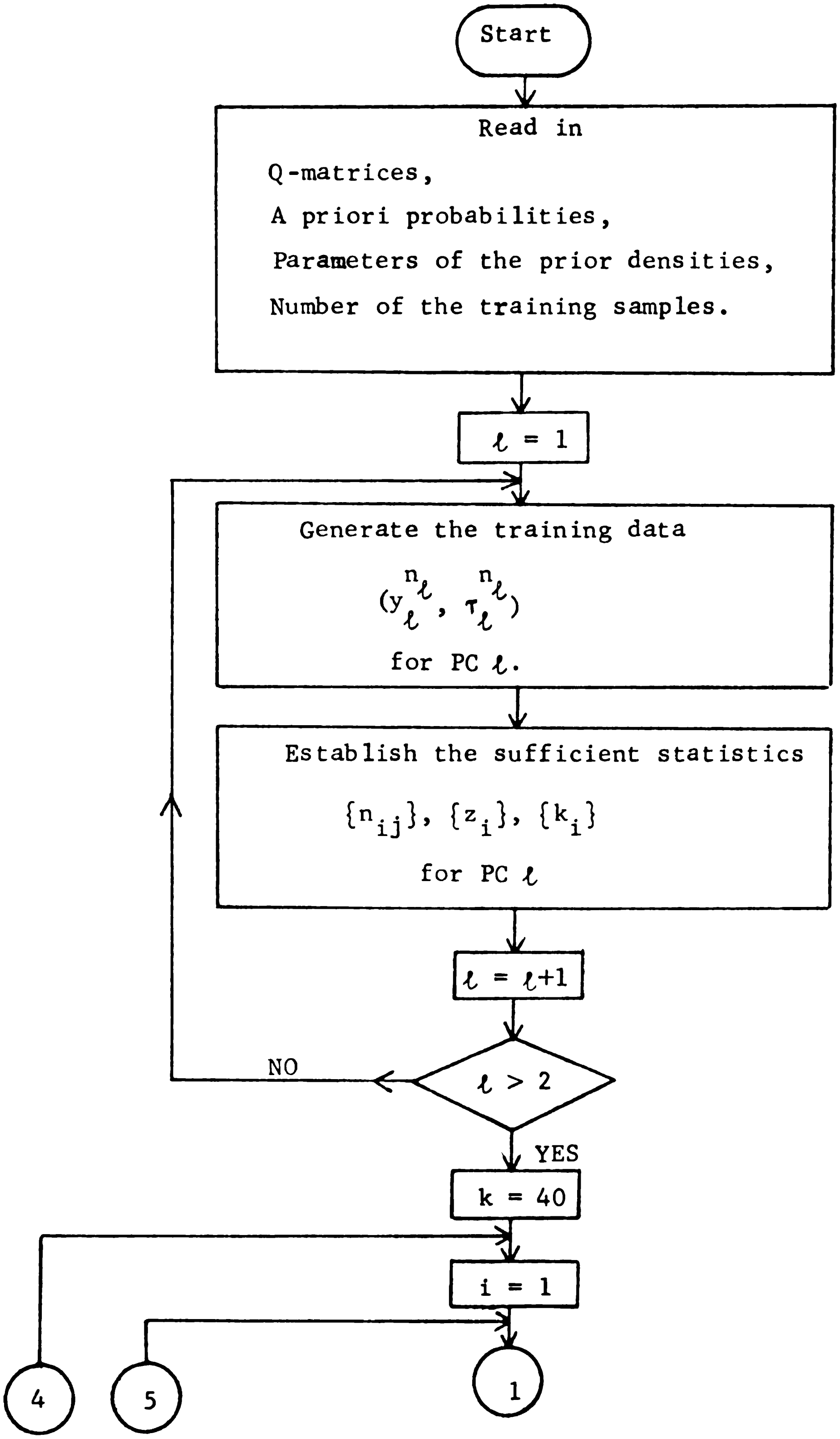
$$= \frac{1}{r^o_{ij}} \delta(\frac{q_{ij} - q^o_i r^o_{ij}}{r^o_{ij}}) .$$

Since, $q^o_i \triangleq q^o_i r^o_{ij}$ and $\delta(\frac{q_{ij} - q^o_{ij}}{r^o_{ij}}) = r^o_{ij}\delta(q_{ij} - q^o_{ij})$, the result follows. That is,

$$\lim_{n\to\infty} \psi(q_{ij}|y^n,\tau^n) = \delta(q_{ij} - q^o_{ij}) \quad \forall \; i,j \in \Lambda \quad (j \neq i).$$

APPENDIX D

ALGORITHMS FOR ERROR CURVES AND FOR DECISION MAKING


Start
Read in
Q-matrices,
A priori probabilities,
Parameters of the prior densities,
Number of the training samples.
$\ell = 1$
Generate the training data
$(y_\ell^{n_\ell}, \tau_\ell^{n_\ell})$
for PC ℓ.
Establish the sufficient statistics
$\{n_{ij}\}, \{z_i\}, \{k_i\}$
for PC ℓ
$\ell = \ell+1$
$\ell > 2$
NO
YES
k = 40
i = 1
4
5
1

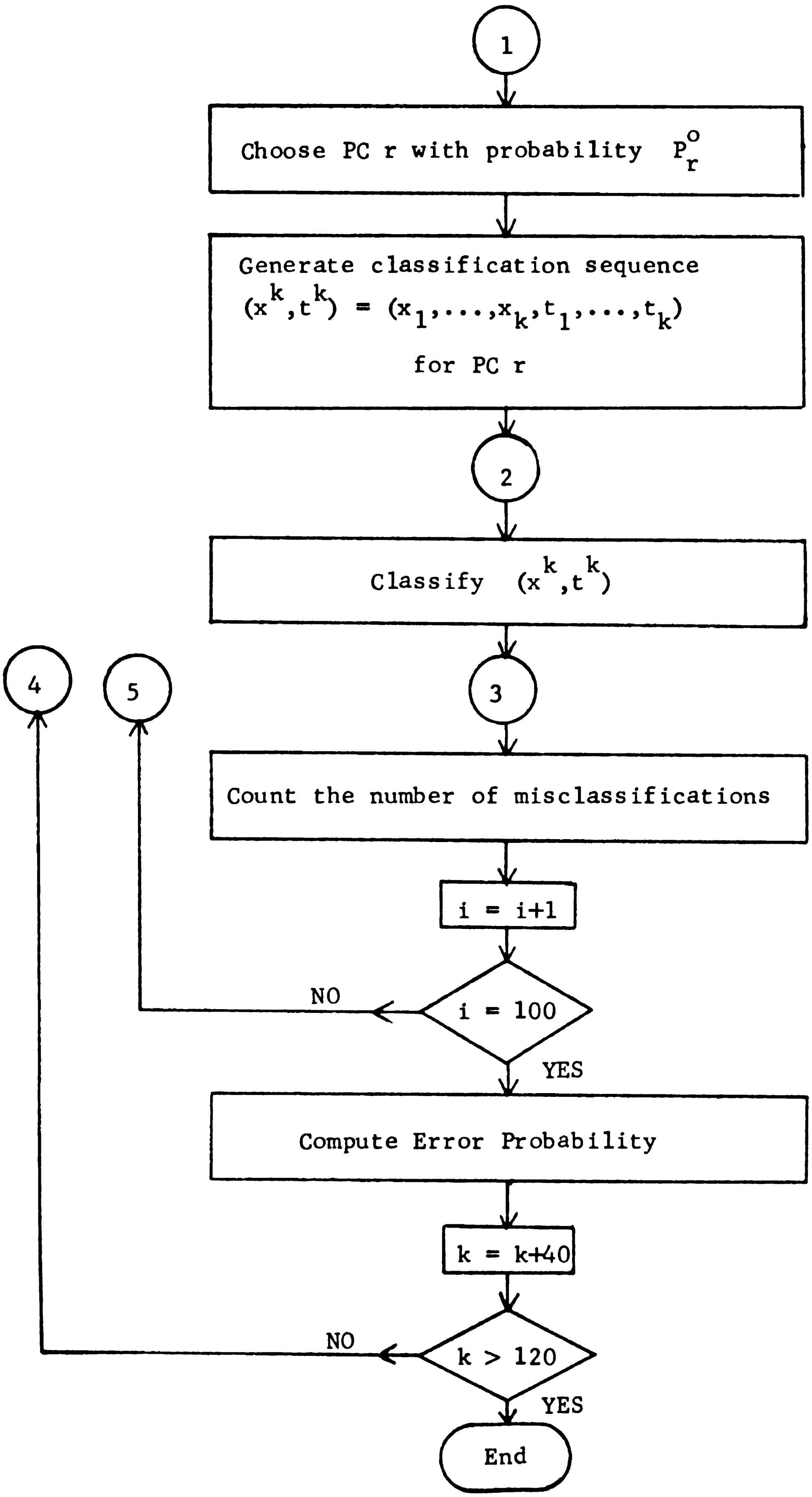
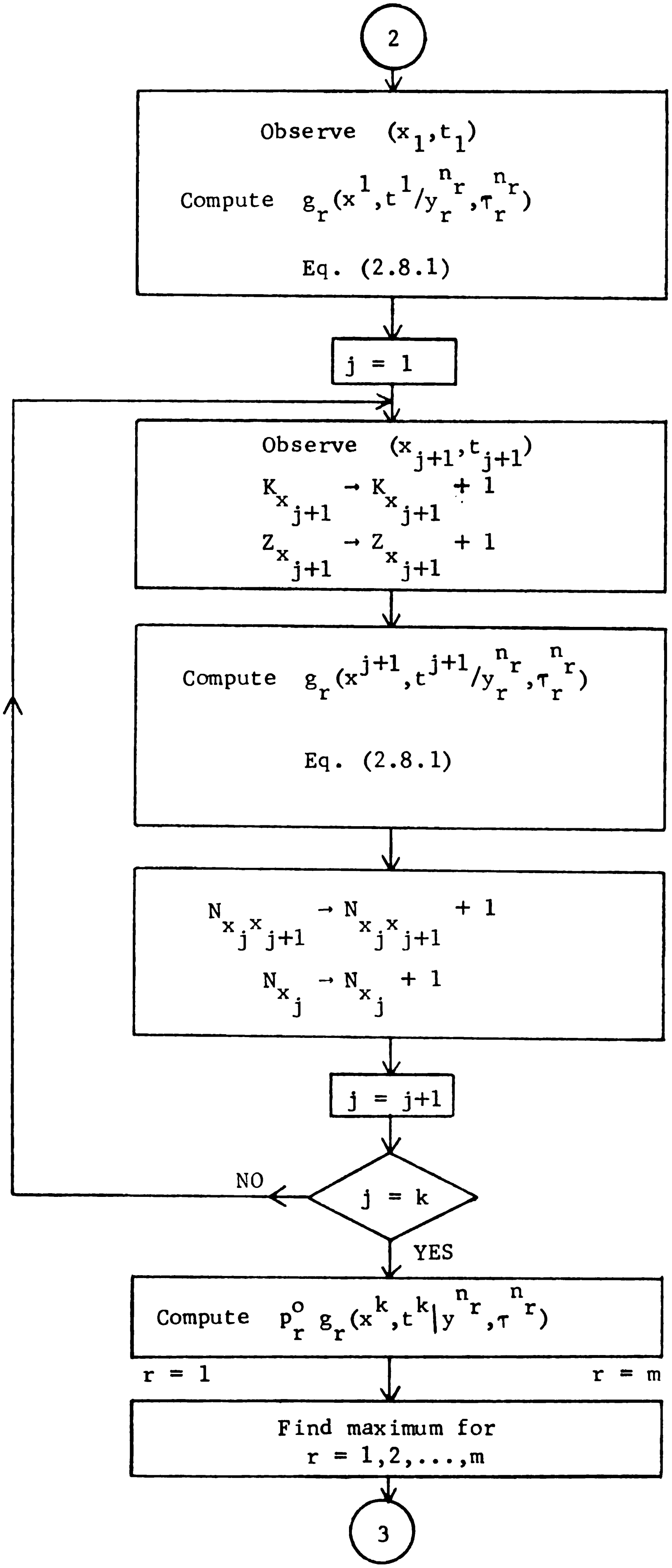

1
Choose PC r with probability P_r^o
Generate classification sequence
$(x^k,t^k) = (x_1,\dots,x_k,t_1,\dots,t_k)$
for PC r
2
Classify (x^k,t^k)
4
5
3
Count the number of misclassifications
i = i+1
NO
i = 100
YES
Compute Error Probability
k = k+40
NO
k > 120
YES
End

Figure D.1: Algorithms for a single error curve and for decision making.


2
Observe (x_1,t_1)
Compute $g_r(x^1,t^1/y_r^{n_r},\tau_r^{n_r})$
Eq. (2.8.1)
$j = 1$
Observe (x_{j+1},t_{j+1})
$K_{x_{j+1}} \rightarrow K_{x_{j+1}} + 1$
$Z_{x_{j+1}} \rightarrow Z_{x_{j+1}} + 1$
Compute $g_r(x^{j+1},t^{j+1}/y_r^{n_r},\tau_r^{n_r})$
Eq. (2.8.1)
$N_{x_j x_{j+1}} \rightarrow N_{x_j x_{j+1}} + 1$
$N_{x_j} \rightarrow N_{x_j} + 1$
$j = j+1$
NO
$j = k$
YES
Compute $p_r^o\, g_r(x^k,t^k|y^{n_r},\tau^{n_r})$
$r = 1$
$r = m$
Find maximum for $r = 1,2,\ldots,m$
3

APPENDIX E

THE PROOF OF LEMMA 3.3.2

First, it is shown that $-\infty < m_{Z_1} < 0$.

$$q_{ij} > p_{ij} \Rightarrow q_i > p_i \quad \forall \quad i \in \Lambda$$

From (3.3.19), (3.3.18), (3.3.9), it follows that

$$m_{Z_1} = \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} p_i \frac{q_{ij}}{q_i} \ell n \frac{p_{ij}}{q_{ij}} + \sum_{i=1}^{N} p_i \frac{q_i - p_i}{q_i}$$

Because $q_{ij} > p_{ij} \Rightarrow \ell n \dfrac{q_{ij}}{p_{ij}} > 0$, m_{Z_1} can be rewritten as

$$m_{Z_1} = \sum_{i=1}^{N} \frac{p_i}{q_i} \left[(q_i - p_i) - \sum_{\substack{j=1 \\ j \neq i}}^{N} q_{ij} \ell n \frac{q_{ij}}{p_{ij}} \right]$$

Using the inequality, $\ell n \dfrac{q_{ij}}{p_{ij}} \geq 1 - \dfrac{p_{ij}}{q_{ij}}$, in the above it follows that

$$\sum_{\substack{j=1 \\ j \neq i}}^{N} q_{ij} \ell n \frac{q_{ij}}{p_{ij}} \geq \sum_{\substack{j=1 \\ j \neq i}}^{N} q_{ij} (1 - \frac{p_{ij}}{q_{ij}}) = \sum_{\substack{j=1 \\ j \neq i}}^{N} (q_{ij} - p_{ij}) = q_i - p_i > 0 \; \forall i \in \Lambda$$

Thus,

$$\sum_{\substack{j=1 \\ j \neq i}}^{N} q_{ij}, \ell n \frac{q_{ij}}{p_{ij}} \geq (q_i - p_i) \quad i \in \Lambda \Rightarrow -\infty < m_{Z_1} < 0$$

Now, it is shown that $0 < m_{Z_2} < +\infty$.

From (3.3.20), (3.3.19) and (3.3.9)

$$m_{Z_2} = \sum_{i=1}^{N} \sum_{\substack{j=1\\ j\neq i}}^{N} P_i \frac{p_{ij}}{p_i} \ln \frac{p_{ij}}{q_{ij}} + \sum_{i=1}^{N} P_i \frac{q_i - p_i}{q_i}$$

Using the same argument as in the first part of the proof, m_{Z_2} can be written as

$$m_{Z_2} = \sum_{i=1}^{N} \frac{P_i}{p_i} \left[(q_i - p_i) - \sum_{\substack{j=1\\ j\neq i}}^{N} p_{ij} \ln \frac{q_{ij}}{p_{ij}} \right]$$

Using the inequality, $\ln \frac{q_{ij}}{p_{ij}} \leq \frac{q_{ij}}{p_{ij}} - 1$, in the above it follows that

$$\sum_{\substack{j=1\\ j\neq i}}^{N} p_{ij} \ln \frac{q_{ij}}{p_{ij}} \leq {}_{\substack{j=1\\ j\neq i}} p_{ij} \left(\frac{q_{ij}}{p_{ij}} - 1\right) = \sum_{\substack{j=1\\ j\neq i}}^{N} (q_{ij} - p_{ij}) = q_i - p_i > 0$$

Thus,

$$\sum_{\substack{j=1\\ j\neq i}}^{N} p_{ij} \ln \frac{q_{ij}}{p_{ij}} \leq (q_i - p_i) \qquad i \in \Lambda \Rightarrow 0 < m_{Z_2} < +\infty$$

and the lemma is proved.