https://ntrs.nasa.gov/search.jsp?R=19700031721 2018-07-23T12:17:15+00:00Z

T- 7001561

70 41037

CR 113864

Office of Naval Research

Contract N00014-67-A-0298-0006 NR-372-012

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

Grant NGL 22-007-143

LEARNING WITH A PROBABILISTIC TEACHER

CASE FILE COPY

By

Ashok K. Agrawala

May 1970

Technical Report No. 611

RECEIVED
NASA STI FACILITY
INPUT BRANCH
OCT 1970

This document has been approved for public release and sale; its distribution is unlimited. Reproduction in whole or in part is permitted by the U. S. Government.

Division of Engineering and Applied Physics
Harvard University • Cambridge, Massachusetts

Office of Naval Research

Contract N00014-67-A-0298-0006

NR-372-012

National Aeronautics and Space Administration

Grant NGL 22-007-143

LEARNING WITH A PROBABILISTIC TEACHER

By

Ashok K. Agrawala

Technical Report No. 611

This document has been approved for public release and sale; its distribution is unlimited. Reproduction in whole or in part is permitted by the U. S. Government.

May 1970

The research reported in this document was made possible through support extended the Division of Engineering and Applied Physics, Harvard University by the U. S. Army Research Office, the U. S. Air Force Office of Scientific Research and the U. S. Office of Naval Research under the Joint Services Electronics Program by Contracts N00014-67-A-0298-0006, 0005, and 0008 and by the National Aeronautics and Space Administration under Grant NGL 22-007-143.

Division of Engineering and Applied Physics

Harvard University · Cambridge, Massachusetts

LEARNING WITH A PROBABILISTIC TEACHER

By

Ashok K. Agrawala

Division of Engineering and Applied Physics

Harvard University · Cambridge, Massachusetts

ABSTRACT

Estimation or learning problems arise in practical systems in many ways. Depending on the learning information available, the estimation problem may be supervised or unsupervised. Bayesian estimation may be used for both these problems. The Bayesian solution of a supervised learning problem is reasonably simple while the unsupervised Bayesian learning is enormously complex. A practical way of solving an unsupervised learning problem is to convert it into a supervised learning problem by labelling the observation before using it for learning. Decision directed learning scheme uses the result of a decision process as the label. The computations for this scheme are feasible but the resulting estimates do not converge to the correct value.

A learning scheme, 'learning with a probabilistic teacher', is proposed in which a label is generated as a random variable from an appropriate probability density function. This scheme leads to a feasible solution to an unsupervised learning problem and assures the convergence of the estimate to the correct value. The average mean square error of the resulting estimate is twice the mean square error of the 'learning

with a teacher' estimate. This learning scheme can also be used to estimate the state of a Gauss Markov sequence when the observation process has additive as well as multiplicative noise.

TABLE OF CONTENTS

		Page
ABSTRACT		i
LIST OF FIGURES		v
LIST OF SYMBOLS		vii
Chapter I	INTRODUCTION	1-1
I. 1.	Estimation Problems in Pattern Classification	1-3
	I. 1. 1. Supervised and Unsupervised Learning in Pattern Classification	1-5
I. 2.	Supervised and Unsupervised Learning	1-5
	I. 2. 1. Problem Formulation	1-6
I. 3.	Bayesian Estimation Philosophy and Technique	1-7
	I. 3. 1. The Computation of the Posterior Density (Batch Processing)	1-8
	I. 3. 2. Definition of the Complexity of Computations Measure	1-10
	I. 3. 3. The Computation of the Posterior Density (Recursive Processing)	1-10
	I. 3. 4. Convergence of Bayesian Learning Scheme	1-13
I. 4.	The Implementation of the Bayesian Learning Scheme	1-13
	I. 4. 1. An Example	1-16
I. 5.	Labelling	1-20
	I. 5. 1. Labelling Method I. Decision Directed Learning	1-23
I. 6.	The Relation of Labelling to the Bayesian Learning Scheme	1-26
I. 7.	Summary	1-28
Appendix A - More about Decision Directed Learning		1-30
References for Chapter I		1-32
Chapter II	LEARNING WITH A PROBABILISTIC TEACHER	2-1
II. 1.	A Class of Unsupervised Learning Problems	2-1
II. 2.	Learning with a Probabilistic Teacher	2-2
	II. 2. 1. The Probabilistic Labelling	2-3
	II. 2. 2. Updating	2-5
	II. 2. 3. The Operation of LPT Scheme	2-7
II. 3.	The Convergence of LPT Scheme	2-8
II. 4.	The Implementation of the LPT Scheme	2-12
II. 5.	Examples	2-15
	II. 5. 1. Example C-1	2-16
	II. 5. 2. Example C-2	2-23

Page

II. 6. The LPT Estimate: Some Properties 2-29
II. 7. Summary 2-32
Appendix A - The Convergence of the LPT Scheme 2-34
Appendix B - An Example Violating the Assumption (A4) 2-41
Appendix C - Block Diagram Structure of Solutions to Learning Problems in Pattern Classification 2-45
References for Chapter II 2-48

Chapter III APPLICATION OF THE LPT SCHEME TO GAUSS MARKOV SEQUENCE 3-1

III. 1. Problem Formulation - Problem D 3-3
III. 2. General Solution 3-5
III. 3. The LPT Solution 3-7
III. 4. The Implementation of the LPT Solution 3-12
III. 5. Example 3-14
III. 5. 1. The Best Linear Filter for Problem (D-1) (Nahi's Solution) 3-16
III. 5. 2. Numerical Results 3-17
III. 6. Summary 3-19
References for Chapter III 3-20

Chapter IV CONCLUDING REMARKS AND SUGGESTIONS FOR FUTURE WORK 4-1

IV. 1. Suggestions for Further Work 4-3

ACKNOWLEDGEMENTS

LIST OF FIGURES

		Page
Fig. 1.1.	Labelled Learning	1-22
Fig. 1.2.	Decision Directed "Learning" Region	1-25
Fig. 2.1.	A ψ Function for Discrete $\mathcal{H}$	2-14
Fig. 2.2.	Two Typical Learning Sequences of $\bar{x}_k$ for Example C-1	2-20
Fig. 2.3.	Sample Variance of $\bar{x}_k$	2-22
Fig. 2.4.	Gamma 2 Densities v = 1	2-25
Fig. 2.5.	Two Typical Learning Sequences of v_k for Problem C-2	2-28
Fig. 2.6.	Sample Variance of v_k	2-30
Fig. 2.7.	Example of Appendix II-B	2-42
Fig. 3.1.	The LPT Solution of Problem D	3-11
Fig. 3.2.	Mean Square Error of Estimates for Problem (D-1)	3-18

LIST OF SYMBOLS

1. A variable with an underline is considered as a random variable e.g. $\underline{z}$.
2. A probability density function for random variable $\underline{x}$ is written as $p_{\underline{x}}$. Its functional dependence on x is shown as $p_{\underline{x}}(x)$. The parameters of this density function may be shown explicitly after a colon in the parentheses e.g. $p_{\underline{x}}(x:H)$.
3. A conditional density function of $\underline{x}$ given $\underline{z} = z$ is written as $p_{\underline{x}/\underline{z}}(x; z)$.
4. A subscript k denotes the kth element of a sequence.
5. A script letter denotes a sequence e.g. $\mathcal{z}_k = z_1, z_2, \ldots, z_k$.

Symbol		Page
E	expectation operator	2-31
E_x	a set belonging to $\mathcal{X}$, the space of x	2-39
f	a function	1-1
F_{1k}	variable defined by Eq. (3.28)	3-16
F_{2k}	variable defined by Eq. (3.29)	3-16
g_k	variable defined by Eq. (A.1)	2-34
H	variable, class, hypothesis, an element of space $\mathcal{H}$	1-2
H^i	i^{th} element of a discrete space $\mathcal{H}$	1-3
$\mathcal{H}$	space on which H is defined	1-3
$\mathcal{h}_k$	sequence $(H_1, H_2, \ldots, H_k)$	2-29
I_{E_x}	indicator function for set E_x	2-39
J	cost function	1-7
k	integer number	1-5
ℓ	label, an element of space $\mathcal{H}$	1-2

Symbol		Page
ℓ^{dd}	a label generated by decision directed scheme	2-7
$\mathcal{L}_k$	sequence $(\ell_1, \ell_2, \ldots, \ell_k)$	1-21
ℓ^i	i^{th} element of a vector ℓ	3-13
L	variable defined by Eq. (2.20)	2-17
M	constant	2-31
M_k	variable defined in Appendix II-B.	2-41
	covariance of a normal distribution	3-4
m_k	variable as defined in Appendix II-B	2-43
n	an integer number	1-26
$N(\bar{x}, P)$	normal distribution with mean $\bar{x}$ and variance P	1-16
$p_{\underline{x}}(x)$	probability density function of random variable $\underline{x}$	1-3
$p_{\underline{x}}(x:H)$	H is some parameter of $p_{\underline{x}}$	1-4
$p_{\underline{x}/\underline{z}}(x; z)$	conditional density function of $\underline{x}$ given $\underline{z} = z$	1-4
p	probability of occurance of a class for examples (A1), (B1), (C-1), (C-2), (D-1) etc.	1-16
P	covariance of a normal distribution	1-17
$P_k(E_x)$	variable as defined by Eq. (A-19)	2-39
Q	covariance of a normal distribution	3-3
R	covariance of a normal distribution	1-16
S	variable defined by Eq. (3.29)	3-16
$\underline{u}$	a random variable	2-39
u	a step function	2-41
v	variable for Gamma 2 distribution	2-24
$\underline{v}$	Gaussian white noise	1-16
$\underline{w}$	Gaussian white noise	3-1

Symbol		Page
$\bar{w}$	mean of a normal distribution	3-3
x	unknown parameter	1-1
x_o	correct value of x	1-7
$\bar{x}$	mean of a normal distribution	2-17
$\hat{x}$	an estimate of x	1-7
$\hat{x}^A$	estimate of x for problem A	2-32
$\mathbb{X}$	space on which x is defined	1-4
y	a classified sample (z, H)	1-3
y'	a labelled sample (z, ℓ)	1-20
$\mathcal{Y}_k$	sequence $(y_1, y_2, \ldots, y_k)$	1-5
$\mathbb{Y}$	space on which y is defined, $\mathbb{Z} \times \mathcal{H}$	1-3
z	observation, sample	1-1
$\mathcal{Z}_k$	sequence $(z_1, z_2, \ldots, z_k)$	1-6
$\mathbb{Z}$	space on which z is defined	1-3
α_k	probability $p_{\underline{H}_k/\underline{z}_{k'}, -k-1', -k-1}$ for examples (C-1), (C-2), etc.	2-18
β_k	as defined in example C-2	2-26
ϕ	loss function	1-7
	transition (matrix) of a gauss markov sequence	3-1
	a function	2-39
μ	mean of a normal distribution	2-23
ν	variable of Gamma 2 distribution	2-24
ψ	a function	2-13
$\underline{\omega}$	a random number from space Ω	2-13
Ω	space on which $\underline{\omega}$ is defined	2-13

Symbol		Page
Γ	variable as defined by Eq. (3.1)	3-1
$\delta(\cdot)$	delta function	1-13
$\delta_{k_1 k_2}$	kroneker delta	3-3

CHAPTER I

INTRODUCTION

An estimation problem arises in practical systems when the value of some parameter x of the system is unknown and some measurement (or observation) z on the system is available. If $z = f(x)$ where f is a known function, the value of x may be obtained from z by solving this implicit equation. However, in many practical systems some parameter v of the function f has to be treated as a random variable. Now the knowledge of the function f is not sufficient to get the value of x. It can only be "estimated" from the observation using some statistical estimation technique. In addition to the knowledge of the function f such techniques require the statistics of the noise $\underline{v}$.

The complexity of computations required by any estimation scheme depends on the form of f and the statistics of $\underline{v}$. A scheme can be used in practice only if the computations for it are feasible. In addition it should give 'good' estimates of x. It should be able to use a sequence of observations if available, and result in an estimate which, in the limit, converges to the correct value of x.

In some estimation problems two parameters of f are random variables. The statistics of the second parameter $\underline{H}$ may also be available. Such problems naturally arise in pattern recognition context where they are referred to as unsupervised estimation (or learning) problems. If, on the other hand, the correct value of $\underline{H}$ is available, the estimation is called supervised learning. The presence of additional noise makes unsupervised learning more complex than supervised learning.

Bayesian estimation may be used to get an optimal solution to these learning problems. It is reasonably simple for the supervised

learning. The solution gets enormously complex for an unsupervised learning problem and becomes infeasible. A practical way of solving an unsupervised learning problem is to convert it into a supervised learning problem by first estimating $\underline{H}$ as ℓ and then treating ℓ (the label) as the correct value of $\underline{H}$. The generation of the label ℓ may be called labelling. Now the labelling, as well as the supervised learning which follows it, should be feasible and result in a converging estimator.

The only scheme proposed in literature which makes use of labelling is 'decision directed' learning scheme. The computations for this scheme are feasible but the resulting estimates do not converge to the correct value. In this work a feasible labelling method is proposed which assures that the estimate converges to the correct value.

The proposed labelling method uses a random variable $\underline{\ell}$ as an estimate of $\underline{H}$. Therefore the resulting learning scheme may be called 'learning with a probabilistic teacher'. This learning scheme is formulated in Chapter II. Its convergence is established. Some examples are presented to show the behavior of the estimates.

In some estimation problems the parameter x does not remain constant over the observation period. The observation process may still make it an unsupervised estimation problem. In Chapter III we show how the proposed learning technique can be used when x varies as a Gauss Markov sequence.

The learning scheme proposed in this work opens up a new line of attack on unsupervised learning problems. Various problems require investigation in this connection. Chapter IV contains some suggestions for further work.*

---

*Unless otherwise specified, all the variables are considered as continuous variables in this work.

I.1. Estimation Problems in Pattern Classification

In a pattern classification (or simply classification) problem, given an object and a set of classes from which the object may have been drawn, we have to determine the class from which the object was drawn. To put it in a mathematical framework, let $\mathcal{Z}$ be the space in which the object z is defined and $\mathcal{H}$ the set of classes (or hypotheses or labels). To solve the pattern classification problem, for any given or observed value z we have to determine $H^i \in \mathcal{H}$, the class from which z was drawn (or the hypothesis which was active when z was drawn). The pattern classification problem, therefore, is to determine a way to process the observation z to make a classification decision which is "optimal" in some sense.

To define the sense in which the classification decision is optimal a loss function associated with the misclassification is given. This loss function depends, both on the correct class of z as well as the class to which it is classified. A reasonable definition of the optimum decision system is the decision system which minimizes the expected loss function. Such a system is nothing but the realization of a Bayes decision rule [1].

When posed this way the pattern classification problem is characterized by the joint density function $p_{\underline{y}}(y) = p_{\underline{z},\underline{H}}(z, H)$ defined over the space $\mathcal{Y} = \mathcal{Z} \times \mathcal{H}$. Depending on the amount of information available three categories of classification problems are possible, that

(a) We have the complete knowledge of $p_{\underline{y}}(y)$,

(b) we have no knowledge of $p_{\underline{y}}(y)$, and

(c) we have partial knowledge of $p_{\underline{y}}(y)$ in terms of the functional

form of $p_{\underline{z},\underline{H}}(z, H^i:x)$. We do not know the values of some parameters $x \in \underline{X}$.

The problems in category (a) can be completely solved using Bayes decision rule [1]. The problems in the second category are commonly referred to as 'non-parametric' decision problems. Reference [2] contains a survey of techniques applicable to the non-parametric problems, among other things.

For the problems of category (c) we know the functional form of the joint probability density function $p_{\underline{z},\underline{H}}(z, H:x)$. If we can treat the unknown parameter x as a random variable, $\underline{x}$, and summarize the uncertainty in our knowledge about it as a prior probability density function $p_{\underline{x}}(x)$, we may express $p_{\underline{z},\underline{H}}(z, H)$ as

$$p_{\underline{z},\underline{H}}(z, H) = \int_{\underline{X}} p_{\underline{z},\underline{H}/\underline{x}}(z, H; x) p_{\underline{x}}(x) dx$$

where

$$p_{\underline{z},\underline{H}/\underline{x}}(z, H; x) = p_{\underline{z},\underline{H}}(z, H:x)$$

Now we can use the Bayesian decision techniques applicable to category (a). If, in addition, a sequence of observations from $(\mathcal{Z} \times \mathcal{H})$ space is available, we may try to use the information contained in this sequence in improving our knowledge about $\underline{x}$. We may do this by estimating the value of $\underline{x}$ using the given sequence.

Therefore, we see that in the pattern classification problems of category (c), an estimation problem arises when we have the additional knowledge available as the sequence of observations. In pattern classification literature such estimation problems are also called 'learning' problems.*

---

*In this work we shall use 'estimation' and 'learning' interchangeably.

I.1.1. Supervised and Unsupervised Learning in Pattern Classification

In a learning problem associated with a pattern classification problem of category (c), we require a sequence of observations from the space $\mathcal{Y}$, which can be used to estimate or learn the value of the unknown parameter. This sequence of observations is also called the 'learning information'. The learning information may be available in one of the two forms:

a) The observed sequence has the form $\mathcal{Y}_k = [y_1, y_2, \ldots, y_k] = [(z_1, H_1), (z_2, H_2), \ldots, (z_k, H_k)]$, i.e. we are given the correct classification for all the observed values of $\underline{z}$ in the sequence.

b) The observed sequence has the form $\mathcal{Z}_k = [z_1, z_2, \ldots, z_k]$. We are given the observed values of $\underline{z}$ in the sequence with no information regarding the classification of each of the z_k's.

The structure of the learning procedure depends on the form in which the learning information is available. When the learning information has the form (a) the correct classification for each observed value of $\underline{z}$ in the sequence is also available. Some supervision of the sequence of observations (by some external means) is required to generate the correct classifications. Hence the learning problem using this information is called 'supervised learning' or 'learning with a teacher' problem. The learning information in the second form does not require any external supervisions and hence is called 'unsupervised learning' or 'learning without a teacher' problem.

I.2. Supervised and Unsupervised Learning

So far we have seen how the supervised and unsupervised estimation problems occur in pattern classification. To consider these

estimation problems formally, in this section we formulate these problems in general. The formulation here will be the basis of all our discussions in this work.

I.2.1. Problem Formulation

We consider the following

1) $\underline{z}$ is a random variable defined in $\mathcal{Z}$ space.*
2) $\underline{H}$ is a random variable defined in $\mathcal{H}$ space, the space of classes or hypotheses.
3) A joint probability density function $p_{\underline{y}}(y)$ is defined on $\mathcal{Y} = (\mathcal{Z} \times \mathcal{H})$ space.
4) A density function $p_{\underline{H}}(H)$ is defined on $\mathcal{H}$ and is known.+

   Therefore we may express $p_{\underline{y}}(y)$ as

$$p_{\underline{y}}(y) = p_{\underline{z},\underline{H}}(z, H) = p_{\underline{z}/\underline{H}}(z; H) p_{\underline{H}}(H) \quad . \tag{1.1}$$

5) The conditional density functions $p_{\underline{z}/\underline{H}}(z; H:x)$ have some unknown parameters x. We are given the functional form of this conditional density function.
6) The correct value of the unknown parameter x is x_o. Based on this structure we define the following two problems:

Problem A - Given a sequence of observations in the form $\mathcal{Y}_k = (\mathcal{Z}_k, \mathcal{H}_k) = [(z_1, H_1), \ldots, (z_k, H_k)]$ where

$$p_{\underline{z}_k, \underline{H}_k}(z_k, H_k) = p_{\underline{z}, \underline{H}}(z_k, H_k) \tag{1.2}$$

we have to make an optimal estimate of x.

---

*The spaces that we consider here are continuous spaces unless otherwise specified.

+When $\mathcal{H}$ is discrete the density function $p_{\underline{H}}$ will be a collection of delta functions which we accept as an admissible density function.

Problem B - Given a sequence of observations in the form

$\mathcal{Z}_k = [z_1, z_2, \ldots, z_k]$ where

$$p_{\underline{H}_k}(H_k) = p_{\underline{H}}(H_k)$$

and

$$p_{\underline{z}_k}(z_k) = p_{\underline{z}}(z_k) = \int_{\mathcal{H}} p_{\underline{z}/\underline{H}}(z_k; H) p_{\underline{H}}(H) dH^* \qquad (1.3)^{**}$$

We have to make an optimal estimate of x.

To specify the sense in which we desire the estimate of x to be optimal, let $\phi(\hat{x} - x)$ be the loss function associated with the value $\hat{x}$, the estimated value of x. We define the cost function J as

$$J = E[\{\phi(\hat{x} - x)\} / \text{learning information}] \quad . \qquad (1.4)$$

We shall call the estimate 'optimal' if it minimizes the cost function J.

We note that Problem A is a problem of 'supervised learning' or 'learning with a teacher' while Problem B is a problem of 'unsupervised learning' or 'learning without a teacher'.

I. 3. Bayesian Estimation Philosophy and Technique

Let us examine the Bayesian estimation techniques for solving the problems A and B.

The Bayesian estimation philosophy assumes that x is a random variable $\underline{x}$ defined in some appropriate space $\mathbb{X}$. It further assumes that a prior distribution $p_{\underline{x}}(x)$ is available which summarizes the uncertainty in x. Now to evaluate the value of the cost function J, say

---

* The probability density function $p_{\underline{z}_k}(z_k)$ defined this way is called a 'mixture'.

** The observed z_k is generated from the space $\mathcal{Y}$ as (z_k, H_k), though we are allowed to observe only z_k.

for a single observation $\underline{z} = z$, all we need is the posterior density function $p_{\underline{x}/\underline{z}}(x; z)$. Hence the computation of the posterior density function is also sufficient to find the estimate of x. The same idea applies when we observe a sequence. Now we have to calculate the posterior density function for this sequence.

The central idea of the Bayesian estimation scheme is the computation of the posterior density function. Therefore, by estimation we shall imply the computation of the posterior density function. As a result we may rewrite the statement of Problem A and B of Section I. 2. 1 as follows:

Problem A - The Supervised Learning Problem

Given a sequence of observations $\mathcal{Y}_k = (y_1, y_2, \ldots, y_k)$ and an a priori density function $p_{\underline{x}}(x)$ compute the posterior density function $p_{\underline{x}/\mathcal{Y}_k}(x; \mathcal{Y}_k)$.

Problem B - The Unsupervised Learning Problem

Given a sequence of observations $\mathcal{Z}_k = (z_1, z_2, \ldots, z_k)$ and an a priori density function $p_{\underline{x}}(x)$ compute the posterior density function $p_{\underline{x}/\mathcal{Z}_k}(x; \mathcal{Z}_k)$.

Let us see how the posterior density function can be computed.

I. 3. 1. The Computation of the Posterior Density (Batch Processing)

Using the Bayes rule we may express the posterior density function for the problem A as

$$p_{\underline{x}/\mathcal{Y}_k}(x; \mathcal{Y}_k) = \frac{p_{\mathcal{Y}_k/\underline{x}}(\mathcal{Y}_k; x)}{p_{\mathcal{Y}_k}(\mathcal{Y}_k)} \; p_{\underline{x}}(x) \tag{1.5}$$

where

$$p_{\underline{\mathcal{Y}}_k}(\mathcal{Y}_k) = \int_{\underline{X}} p_{\underline{\mathcal{Y}}_k/\underline{x}}(\mathcal{Y}_k; x) p_{\underline{x}}(x) dx \tag{1.6}$$

Here

$$p_{\underline{y}/\underline{x}}(y; x) = p_{\underline{z}, \underline{H}/\underline{x}}(z, H; x) = p_{\underline{z}/\underline{H}}(z; H:x) p_{\underline{H}}(H) \quad . \tag{1.7}$$

Knowing the values for $\mathcal{Y}_k$ we can use these equations to compute the posterior density function $p_{\underline{x}/\underline{\mathcal{Y}}_k}(x; \mathcal{Y}_k)$.

The Eq. (1.5) requires the knowledge of $p_{\underline{\mathcal{Y}}_k/\underline{x}}(\mathcal{Y}_k; x)$ which is the joint conditional density $p_{\underline{y}_1, \underline{y}_2, \cdots \underline{y}_k/\underline{x}}(y_1, y_2, \cdots, y_k; x)$. Let us assume that

$$\text{(A1)} \qquad p_{\underline{\mathcal{Y}}_k/\underline{x}}(\mathcal{Y}_k; x) = \prod_{i=1}^{k} p_{\underline{y}_i/\underline{x}}(y_i; x)^* \tag{1.8}$$

i.e. $\underline{y}_i$'s are conditionally independent, given $\underline{x}$. This assumption leads to some simplifications in the equations above.

A very similar computation is required to compute the posterior density function $p_{\underline{x}/\underline{\mathcal{Z}}_k}(x; \mathcal{Z}_k)$ for Problem B. We may express $p_{\underline{x}/\underline{\mathcal{Z}}_k}(x; \mathcal{Z}_k)$ as

$$p_{\underline{x}/\underline{\mathcal{Z}}_k}(x; \mathcal{Z}_k) = \frac{p_{\underline{\mathcal{Z}}_k/\underline{x}}(\mathcal{Z}_k; x)}{p_{\underline{\mathcal{Z}}_k}(\mathcal{Z}_k)} p_{\underline{x}}(x) \tag{1.9}$$

where

---

* Under this assumption the conditional independence of $\underline{z}_i$'s, given $\underline{x}$, follows by integrating both sides of (1.8) over the space on which the sequence $\mathcal{H}_k$ is defined. This implies

$$p_{\underline{\mathcal{Z}}_k/\underline{x}}(\mathcal{Z}_k; x) = \prod_{i=1}^{k} p_{\underline{z}_i/\underline{x}}(z_i; x) \quad . \tag{1.8a}$$

$$p_{\mathfrak{z}_k}(\mathfrak{z}_k) = \int_{\underline{X}} p_{\mathfrak{z}_k/\underline{x}}(\mathfrak{z}_k; x) p_{\underline{x}}(x) dx \tag{1.10}$$

and

$$p_{\mathfrak{z}_k/\underline{x}}(\mathfrak{z}_k; x) = \int_{\mathcal{H}} \int_{\mathcal{H}} \cdots \int_{\mathcal{H}} p_{\mathfrak{z}_k, \underline{h}_k/\underline{x}}(\mathfrak{z}_k, \mathfrak{h}_k; x) dH_1, dH_2, \ldots, dH_k \tag{1.11}$$

The forms of Eqs. (1.9) and (1.10) are very similar to the forms of (1.5) and (1.6) respectively. For Problem B we require the additional (1.11) to take into account the uncertainty about the classifications of the observed sequence.

The computations for the posterior density function using (1.5), (1.6) and (1.7) for Problem A, or (1.9), (1.10) and (1.11) for Problem B require observing the k elements of the sequence. This arrangement of computations may, therefore, be called the Batch Processing mode.

I.3.2. Definition of the Complexity of Computations Measure

A meaningful measure of the complexity of computations is required to compare the ease of implementation of various schemes. For this purpose, as its measure of complexity, we shall use the number of words of computer storage required for any scheme or computation. Here we assume that a number can be stored in one word of computer memory with arbitrary degree of accuracy.

I.3.3. The Computation of the Posterior Density (Recursive Processing)

As we noted in Section I.3.1 the batch processing mode operates on the complete sequence of observations. If the observations are made sequentially, we have to store all k observations before we can start

the computations. This adds to the computational complexity of the batch processing mode. We may avoid this additional complexity by arranging the computations in a recursive manner, such that we compute the posterior density function after each new observation from the sequence.

The computations can be made recursive by arranging them as follows. For Problem A we express the posterior density function $p_{\underline{x}/\underline{y}_k}(x;\underline{y}_k)$ as

$$p_{\underline{x}/\underline{y}_k}(x;\underline{y}_k) = \frac{p_{\underline{y}_k/\underline{x},\underline{y}_{k-1}}(y_k;x,\underline{y}_{k-1})}{p_{\underline{y}_k/\underline{y}_{k-1}}(y_k;\underline{y}_{k-1})}\, p_{\underline{x}/\underline{y}_{k-1}}(x;\underline{y}_{k-1}) \tag{1.12}$$

where

$$p_{\underline{y}_k/\underline{y}_{k-1}}(y_k;\underline{y}_{k-1}) = \int_{\underline{X}} p_{\underline{y}_k/\underline{x},\underline{y}_{k-1}}(y_k;x,\underline{y}_{k-1})\, p_{\underline{x}/\underline{y}_{k-1}}(x;\underline{y}_{k-1})\,dx \tag{1.13}$$

Under the assumption (A1) of conditional independence, these equations can be simplified as

$$p_{\underline{x}/\underline{y}_k}(x;\underline{y}_k) = \frac{p_{\underline{y}_k/\underline{x}}(y_k;x)}{p_{\underline{y}_k/\underline{y}_{k-1}}(y_k;\underline{y}_{k-1})}\, p_{\underline{x}/\underline{y}_{k-1}}(x;\underline{y}_{k-1}) \tag{1.14}$$

where

$$p_{\underline{y}_k/\underline{y}_{k-1}}(y_k;\underline{y}_{k-1}) = \int_{\underline{X}} p_{\underline{y}_k/\underline{x}}(y_k;x)\, p_{\underline{x}/\underline{y}_{k-1}}(x;\underline{y}_{k-1})\,dx \tag{1.15}$$

Similarly for Problem B we write

$$p_{\underline{x}/\underline{z}_k}(x;\underline{z}_k) = \frac{p_{\underline{z}_k/\underline{x},\underline{z}_{k-1}}(z_k;x,\underline{z}_{k-1})}{p_{\underline{z}_k/\underline{z}_{k-1}}(z_k;\underline{z}_{k-1})}\, p_{\underline{x}/\underline{z}_{k-1}}(x;\underline{z}_{k-1}) \tag{1.16}$$

where

$$p_{\underline{z}_k/\mathcal{Z}_{k-1}}(z_k;\mathcal{Z}_{k-1}) = \int_{\underline{X}} p_{\underline{z}_k/\underline{x},\mathcal{Z}_{k-1}}(z_k;x,\mathcal{Z}_{k-1})\, p_{\underline{x}/\mathcal{Z}_{k-1}}(x;\mathcal{Z}_{k-1})\,dx \tag{1.17}$$

and

$$p_{\underline{z}_k/\underline{x},\mathcal{Z}_{k-1}}(z_k;x,\mathcal{Z}_{k-1}) = \int_{\mathcal{H}} p_{\underline{z}_k/\underline{x},\underline{H}_k,\mathcal{Z}_{k-1}}(z_k;x,H_k,\mathcal{Z}_{k-1}) \cdot p_{\underline{H}_k/\underline{x},\mathcal{Z}_{k-1}}(H_k,x,\mathcal{Z}_{k-1})\,dH_k \quad . \tag{1.18}$$

Let us assume that

(A2) $$p_{\underline{H}_k/\underline{x},\mathcal{Z}_{k-1}}(H_k;x,\mathcal{Z}_{k-1}) = p_{\underline{H}_k}(H_k) = p_{\underline{H}}(H_k) \tag{1.19}$$

i.e. $\underline{H}_k$ is independent of $\underline{x}$ and $\mathcal{Z}_{k-1}$ and is identically distributed. Under this assumption we can simplify (1.18) as

$$p_{\underline{z}_k/\underline{x},\mathcal{Z}_{k-1}}(z_k;x,\mathcal{Z}_{k-1}) = \int_{\mathcal{H}} p_{\underline{z}_k/\underline{x},\underline{H}_k,\mathcal{Z}_{k-1}}(z_k;x,H_k,\mathcal{Z}_{k-1}) \cdot p_{\underline{H}}(H_k)\,dH_k$$

The starting point for both the problems is the a priori density function $p_{\underline{x}}(x)$ which is assumed given or known. The process of computing the posterior density function for the kth step (i.e. a computation of (1.12) for Problem A or (1.16) for Problem B) is known as 'updating'.

The computation of the posterior density function this way, is called 'Bayesian Estimation Scheme' or 'Bayesian Learning Scheme'. Eqs. (1.12), (1.13) or (1.14), (1.15) define a recursive Bayesian Learning Scheme for the supervised learning problem (Problem A) and

Eqs. (1.16), (1.17) and (1.18) define a recursive Bayesian Learning Scheme for the unsupervised learning problem (Problem B).

I.3.4. Convergence of Bayesian Learning Scheme

As formulated above, the use of the Bayesian learning scheme requires computing a sequence of posterior density functions. By the convergence of the Bayesian learning scheme we mean that the sequence of the posterior density functions converges with probability one, to a delta function at the correct value of the unknown parameter x, i.e.

$$\lim_{k\to\infty} p_{\underline{x}/\underline{y}_k}(x;\underline{y}_k) = \delta(x - x_o) \qquad \text{w.p. } 1$$

or

$$\lim_{k\to\infty} p_{\underline{x}/\underline{z}_k}(x;\underline{z}_k) = \delta(x - x_o) \qquad \text{w.p. } 1 \tag{1.20}$$

where x_o is the correct value of the parameter x.

The convergence of the Bayesian learning scheme in this sense has been studied by various authors [3], [4], [5]. The convergence in the form of (1.20) has been established under very general conditions.

In this work we shall consider the convergence only in the sense of (1.20).

I.4. The Implementation of the Bayesian Learning Scheme

For the supervised and the unsupervised learning problem we can define a Bayesian learning scheme. The convergence of this scheme is also guaranteed. Let us examine various questions relating to the practical implementation of this learning scheme.

In principle we can always use the Bayesian learning scheme for the supervised as well as the unsupervised learning problem. All the operations required by the Eqs. (1.14), (1.15), (1.16), (1.17) and (1.18)

are well defined. This, however, calls for the ability to store, and to operate on general continuous functions. This capability is beyond the reach of present day digital computers.

If we consider discrete $\mathcal{X}$, $\mathcal{H}$ and $\mathcal{Z}$ spaces, we may be able to store the complete functions by storing the value of the functions at every point in the space on which they are defined. This method has been studied by Fralick [4]. He suggests some computational procedures for such implementations. Unless the spaces involved contain very few points, this implementation method is very complex. This complexity depends on the size of the spaces involved.

For continuous space we have to deal with continuous functions. The only way of handling such functions in a digital computer is when the functions have a parametric form. Then we can generate the value of the function at any point in this space, knowing the parametric form and the values of the parameters.

In Problem A we use (1.14) and (1.15) for updating. Hence we require the knowledge of the density functions $p_{\underline{y}_k/\underline{x}}$ and $p_{\underline{x}/\underline{\mathcal{Y}}_{k-1}}$ along with the value y_k to compute the posterior density function $p_{\underline{x}/\underline{\mathcal{Y}}_k}$. If we assume that

(A3) $p_{\underline{y}_k/\underline{x}}(y_k; x)$ has a parametric form and $p_{\underline{y}_k/\underline{x}}(y_k; x) = p_{\underline{y}/\underline{x}}(y_k; x)$

i.e. the form of this function remains the same for all k, we only need to store the parameters for $p_{\underline{y}/\underline{x}}$ to generate this function and use it in Eqs. (1.14) and (1.15). Now, if $p_{\underline{x}/\underline{\mathcal{Y}}_k}(x; \mathcal{Y}_k)$ also entertains a parametric form and this parametric form remains the same for all k, we can carry out the required updating for any number of steps. The updating at any stage requires computing the values of the parameters of $p_{\underline{x}/\underline{\mathcal{Y}}_k}$.

When the functional form of the posterior density functions $p_{\underline{x}/\underline{y}_k}$ remains the same for all k, a fixed finite dimensional sufficient statistic exists for these functions [6]. The density functions satisfying this requirement (of fixed form) are called reproducing density functions. Various authors [6], [7] have studied the problems which entertain the reproducing density functions and the conditions under which such density functions exist. Spragins [6] has found that the existence of the reproducing density functions requires assumption (A3) and depends only on the form of $p_{\underline{y}/\underline{x}}$.

For the unsupervised learning problem (Problem B) we have to use Eqs. (1.16), (1.17) and (1.18). The form of (1.16) is like that of (1.14) and the form of (1.17) like (1.15). The density function $p_{\underline{z}_k/\underline{x},\underline{z}_{k-1}}(z_k; x, \underline{z}_{k-1})$ has to be computed using (1.18). The existence of the reproducing density functions for this problem depends on the form of $p_{\underline{z}_k/\underline{x},\underline{z}_{k-1}}$ and hence on the form in which (1.18) effects the computations. In practice <u>no reproducing densities are known to exist for unsupervised learning problems</u>.* So, if we start with a $p_{\underline{x}}(x)$ having a parametric form, and go through the recursive computations for the posterior density function, either the posterior density function will have no parametric form or the number of parameters required for its form will increase exponentially. In either case the computations become extremely complex and infeasible. Therefore the Bayesian learning scheme cannot be used for the unsupervised learning problems with continuous functions.

To elaborate further on these points let us consider an example.

---

*Note added in proof: However K. Prabhu and the Author recently discovered an example of reproducing density under unsupervised learning. We intend to publish this seperately [15].

I.4.1. An Example

Let

$$\underline{z}_k = \underline{H}_k x + \underline{v}_k \tag{1.21}$$

where

$$\underline{H}_k = \begin{cases} H^o & \text{with probability } p \\ 0 & \text{with probability } (1 - p) \end{cases}$$

x is the unknown constant,

$\underline{v}_k$ is a purely random sequence, $p_{\underline{v}_k}(v_k) = p_{\underline{v}}(v_k) \sim N(0, R)$

$\underline{v}_k$ and $\underline{H}_k$ are independent. Depending on the observed sequence, we define the following two problems:

Problem A1 - Observing the sequence $\mathcal{Y}_k = [(z_1, H_1), \ldots, (z_k, H_k)]$ compute the sequence of the posterior density functions $p_{\underline{x}/\mathcal{Y}_k}$.

Problem B1 - Observing the sequence $\mathcal{Z}_k = [z_1, z_2, \ldots, z_k]$ compute the sequence of the posterior density functions $p_{\underline{x}/\mathcal{Z}_k}$.

For this example

$$p_{\underline{H}_k}(H_k) = p_{\underline{H}}(H_k) = p\delta(H_k - H^o) + (1 - p)\delta(H_k) \tag{1.22}$$

$$p_{\underline{z}_k/\underline{H}_k, \underline{x}}(z_k; H^o, x) = \frac{1}{\sqrt{2\pi R}} e^{-\frac{1}{2R}[z_k - H^o x]^2} \tag{1.23}$$

$$p_{\underline{z}_k/\underline{H}_k, \underline{x}}(z_k; 0, x) = \frac{1}{\sqrt{2\pi R}} e^{-\frac{1}{2R}[z_k]^2} \tag{1.24}$$

Let us assume that the prior density function $p_{\underline{x}}$ is given as $N(\bar{x}, P_o)$. Let us see the computation steps for the posterior density function for both these problems.

<u>Solution of Problem A1.</u>

For the observation $\underline{y}_1 = y_1 = (z_1, H_1)$

$$p_{\underline{y}_1}(y_1) = \int_{-\infty}^{\infty} p_{\underline{y}_1/\underline{x}}(y_1; x) p_{\underline{x}}(x) dx$$

$$= \int_{-\infty}^{\infty} \frac{p_{\underline{H}}(H_1)}{\sqrt{2\pi R}} e^{-\frac{1}{2R}(z_1 - H_1 x)^2} \frac{1}{\sqrt{2\pi P_o}} e^{-\frac{1}{2R}(x-\bar{x})^2} dx$$

$$= \frac{p_{\underline{H}}(H_1)}{\sqrt{2\pi\,(R + H_1 P_o H_1^T)}} e^{-\frac{1}{2(R+H_1 P_o H_1^T)}(z_1 - H_1\bar{x})^2}$$

$$p_{\underline{x}/\underline{y}_1}(x; y_1) = \frac{\frac{1}{\sqrt{2\pi R}} e^{-\frac{1}{2R}(z_1 - H_1 x)^2} p_{\underline{H}}(H_1) \frac{1}{\sqrt{2\pi P_o}} e^{-\frac{1}{2P_o}(x-\bar{x}_o)^2}}{\frac{p_{\underline{H}}(H_1)}{\sqrt{2\pi(R + H_1 P_o H_1^T)}} e^{-\frac{(z_1 - H_1\bar{x})^2}{2(R+H_1 P_o H_1^T)}}}$$

or

$$p_{\underline{x}/\underline{y}_1} = \begin{cases} \frac{1}{\sqrt{2\pi P_o}} e^{-\frac{1}{2P_o}(x-\bar{x})^2} & \text{if } H_1 = 0 \\ \frac{1}{\sqrt{2\pi P_1}} e^{-\frac{1}{2P_1}(x-\bar{x}_1)^2} & \text{if } H_1 = H^o \end{cases} \tag{1.25}$$

where

$$\bar{x}_1 = \bar{x} + P_1 H^{oT} R^{-1} (z_1 - H^o \bar{x})$$

$$P_1^{-1} = P_o^{-1} + H^{oT} R^{-1} H^o \tag{1.26}$$

We note that the form of the posterior as well as the prior density functions is gaussian. As a result, the gaussian density function is a reproducing density function for this problem. The mean and the variance of this density function form the sufficient statistic and can be updated using (1.25) and (1.26). This updating can be easily carried out for any number of stages.

Eq. (1.26) defines the well known Kalman filter [8] for this problem. We use the Kalman filter to change the mean and the variance of the posterior density function when the 'teacher' tells us that $H_k = H^o$. Otherwise the posterior density function remains the same as the prior density function for that stage.

Solution of Problem B1.

As $\underline{H}_k$ is a random variable we have to use the Eqs. (1.16), (1.17) and (1.18) for this problem. Using (1.18) we may express $p_{\underline{z}_k/\underline{x}, \mathfrak{z}_{k-1}}$ as

$$p_{\underline{z}_k/\underline{x}, \mathfrak{z}_{k-1}}(z_k; x, \mathfrak{z}_{k-1}) = p_{\underline{z}_k/\underline{x}}(z_k; x)$$

$$= \frac{1}{\sqrt{2\pi R}} [p e^{-\frac{1}{2R}(z_k - H^o x)^2} + (1-p) e^{-\frac{z_k^2}{2R}}] \ . \tag{1.27}$$

And from (1.17) for the first observation z_1

$$p_{\underline{z}_1}(z_1) = \int_{-\infty}^{\infty} p_{\underline{z}_1/\underline{x}}(z_1; x) p_{\underline{x}}(x) dx$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi R}} [p e^{-\frac{1}{2R}(z_k - H^o x)^2} + (1-p) e^{-\frac{z_1^2}{2R}}]$$

$$\cdot \frac{1}{\sqrt{2\pi P_o}} e^{-\frac{1}{2P_o}(x-\bar{x})^2} dx$$

$$= \frac{p}{\sqrt{2\pi(R+H^o P_o H^{oT})}} e^{-\frac{(z_1 - H^o\bar{x})^2}{2(R+H^o P_o H^{oT})}} + \frac{(1-p)}{\sqrt{2\pi R}} e^{-\frac{z^2}{2R}} .$$

Therefore

$$p_{\underline{x}/\underline{z}_1}(x; z_1) = \frac{\frac{1}{\sqrt{2\pi R}} \{ p e^{-\frac{1}{2R}(z_1 - H^o x)^2} + (1-p) e^{-\frac{1}{2R}(z_1)^2} \}}{\frac{p}{\sqrt{2\pi(R + H^o P_o H^{oT})}} e^{-\frac{(z_1 - H^o\bar{x})^2}{2(R+H^o P_o H^{oT})}} + \frac{(1-p)}{\sqrt{2\pi R}} e^{-\frac{z_1^2}{2R}}}$$

$$\cdot \frac{1}{\sqrt{2\pi P_o}} e^{-\frac{1}{2P_o}(x-\bar{x})^2} \qquad (1.28)$$

The form of the posterior density function $p_{\underline{x}/\underline{z}_1}(x; z_1)$ as expressed by (1.28) is not gaussian any more. It is a sum of two gaussian functions and hence is a bimodal function. For the next observation we have to use this as the prior density and compute the posterior density

which, now, will have 4 modes. In general, after k stages of recursive computation the posterior density function will be a weighted sum of 2^k gaussian functions. Three parameters have to be stored for each such gaussian function to reconstruct the posterior density function, i.e. the mean, variance and the weight. Therefore, after k stages the posterior density function requires 3×2^k words of storage. This storage requirement keeps on increasing exponentially. As a result we cannot carry out this computation for more than a few stages and cannot use this method for solving this problem.

I. 5. Labelling

According to the formulation of the Section I. 2. 1 the only difference between Problem A and Problem B is in terms of the learning information. For Problem A the learning information is available as a sequence of 'classified'* samples, $\mathcal{Y}_k = [(z_1, H_1), (z_2, H_2), \ldots, (z_k, H_k)]$, while for Problem B we only have a sequence of unclassified samples, $\mathcal{Z}_k = [z_1, z_2, \ldots, z_k]$. In the solution of Problem A we have the convenience of using the reproducing density functions, while the solution of Problem B requires the ability to manipulate general functions. If we can convert the Problem B into an appropriate Problem A, we may be able to use the computational ease of reproducing densities. To convert the Problem B into such a Problem A, therefore, we should get a "label" ℓ_k from the space $\mathcal{H}$ for each z_k. Now we can treat $(z_k, \ell_k) = y_k'$ as a classified sample.

---

*By a classified sample we mean the value of z_k along with the correct value of H_k which was active for the kth observation.

If we adopt this philosophy of labelling in solving the unsupervised learning problem the solution can be carried out in two steps (Figure 1).

Step I. Labelling - Having observed the sample z_k and knowing the prior density function at this stage as $p_{\underline{x}/\mathfrak{z}_{k-1}}(z; \mathfrak{z}_{k-1}:\mathcal{L}_{k-1})$, where $\mathcal{L}_{k-1} = \ell_1, \ell_2, \ldots \ell_{k-1}$ (the sequence of labels generated so far), a label ℓ_k has to be generated.

Step II. Updating - Using $y_k' = (z_k, \ell_k)$, the prior density function $p_{\underline{x}/\mathfrak{z}_{k-1}}(x; \mathfrak{z}_{k-1}:\mathcal{L}_{k-1})$ has to be updated as the posterior density function $p_{\underline{x}/\mathfrak{z}_k}(x; \mathfrak{z}_k:\mathcal{L}_k)$.

The second step here requires similar computations as the solution of a supervised learning problem. Using Eqs. (1.14) and (1.15) this step can be carried out easily if reproducing densities exist. The solution of the unsupervised learning problem arranged this way, critically depends on the method used in generating the label ℓ_k in step I. We would like the labelling process to be such that

(i) it is computationally feasible i.e. the computations required for it can be carried out in practice, and

(ii) it leads to a sequence of posterior densities which converge in the sense of (1.20).

The sequence of samples for any learning problem (A or B) is generated from a joint density function defined on the space $\mathcal{Y} = (\mathcal{Z} \times \mathcal{H})$. In an unsupervised learning problem we are allowed to observe the value of $\underline{z}_k$ only, though it has some value of $\underline{H}_k$ associated with it as its correct classification. Therefore, there always exists a sequence of labels (the correct classifications) satisfying the second requirement

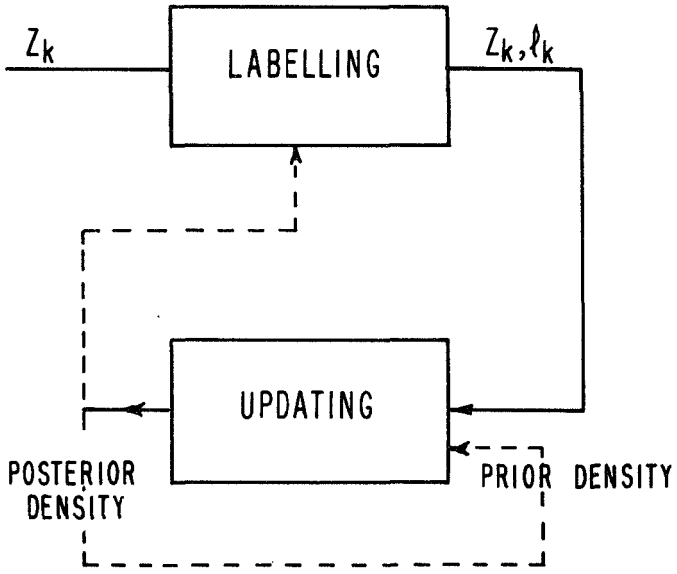


FIG. 1.1 LABELLED LEARNING

above. The labelling here is like a "teacher" that generates a label for the classification of the observed sample. This label is then used as the correct classification in learning.

In selecting the label ℓ_k we can make use of z_k and the prior density function $p_{\underline{x}/\mathcal{Z}_{k-1}}(x; \mathcal{Z}_{k-1}:\mathcal{L}_{k-1})$. One obvious approach of generating the label is by using a decision method for classifying the sample z_k and using this classification as the label. Let us examine this method of labelling.

I. 5. 1. Labelling Method I. Decision Directed Learning

If $\mathcal{H}$ space is discrete we may classify the observed sample z_k with a minimum probability of error and use this classification as the label. To do this we require the values of $p_{\underline{H}_k/\underline{z}_k, \mathcal{Z}_{k-1}}(H_k^i; z_k, \mathcal{Z}_{k-1}:\mathcal{L}_{k-1})$. We may write

$$p_{\underline{H}_k/\underline{z}_k, \mathcal{Z}_{k-1}}(H_k^i; z_k, \mathcal{Z}_{k-1}:\mathcal{L}_{k-1}) =$$

$$\frac{p_{\underline{z}_k/\underline{H}_k, \mathcal{Z}_{k-1}}(z_k; H_k^i, \mathcal{Z}_{k-1}:\mathcal{L}_{k-1}) p_{\underline{H}_k}(H_k^i)}{\sum_i p_{\underline{z}_k/\underline{H}_k, \mathcal{Z}_{k-1}}(z_k; H_k^i, \mathcal{Z}_{k-1}:\mathcal{L}_{k-1}) p_{\underline{H}_k}(H_k^i)} \qquad (1.29)$$

where using assumption (A2) we may write $p_{\underline{z}_k/\underline{H}_k, \mathcal{Z}_{k-1}}(z_k; H_k^i, \mathcal{Z}_{k-1}:\mathcal{L}_{k-1})$ as

$$p_{\underline{z}_k/\underline{H}_k, \mathcal{Z}_{k-1}}(z_k; H_k^i, \mathcal{Z}_{k-1}:\mathcal{L}_{k-1}) =$$

$$\int_{\mathcal{X}} p_{\underline{z}_k/\underline{H}_k, \underline{x}, \mathcal{Z}_{k-1}}(z_k; H_k^i, x, \mathcal{Z}_{k-1}) p_{\underline{x}/\mathcal{Z}_{k-1}}(x; \mathcal{Z}_{k-1}:\mathcal{L}_{k-1}) dx$$

Under assumption (A1),

$$p_{\underline{z}_k/\underline{H}_k,\mathcal{Z}_{k-1}}(z_k; H_k^i, \mathcal{Z}_{k-1}:\mathcal{L}_{k-1}) =$$

$$\int_{\underline{X}} p_{\underline{z}_k/\underline{H}_k,\underline{x}}(z_k; H_k^i, x) p_{\underline{x}/\mathcal{Z}_{k-1}}(x; \mathcal{Z}_{k-1}:\mathcal{L}_{k-1}) dx \quad . \tag{1.30}$$

Next, we generate the label as

$$\ell_k = H_k^i \quad \text{such that} \quad p_{\underline{H}_k/\underline{z}_k,\mathcal{Z}_{k-1}}(H_k^i; z_k, \mathcal{Z}_{k-1}:\mathcal{L}_{k-1}) >$$

$$p_{\underline{H}_k/\underline{z}_k,\mathcal{Z}_{k-1}}(H_k^j; z_k, \mathcal{Z}_{k-1}:\mathcal{L}_{k-1})$$

$$\text{for all } j \neq i \quad . \tag{1.31}$$

This labelling process, therefore, requires the solution of a decision problem and uses the result of the decision as the label. The learning scheme using this method of labelling is called 'Decision Directed Learning'. Scudder [9] has analysed the behavior of this learning scheme.

When we classify the observation using (1.31) we divide the $\mathcal{Z}$ space in various finite (or semi-infinite) regions. Each region has a particular value H^i associated with it. The observed z_k is given the label according to the region in which it lies. Therefore, even though $p_{\underline{H}/\underline{z}}(H; z)$ has a non-zero value for many points of $\mathcal{H}$ space, we label z_k as H_k^i only if it falls in the region of $\mathcal{Z}$ space having this label. As a simple example let us consider the example of Section I. 4. 1 again. If we use the labelling procedure described above we shall be labelling all samples $z_k \geqslant z_o$ as $\ell_k = H^o$ (Figure 2). Therefore, in learning, we

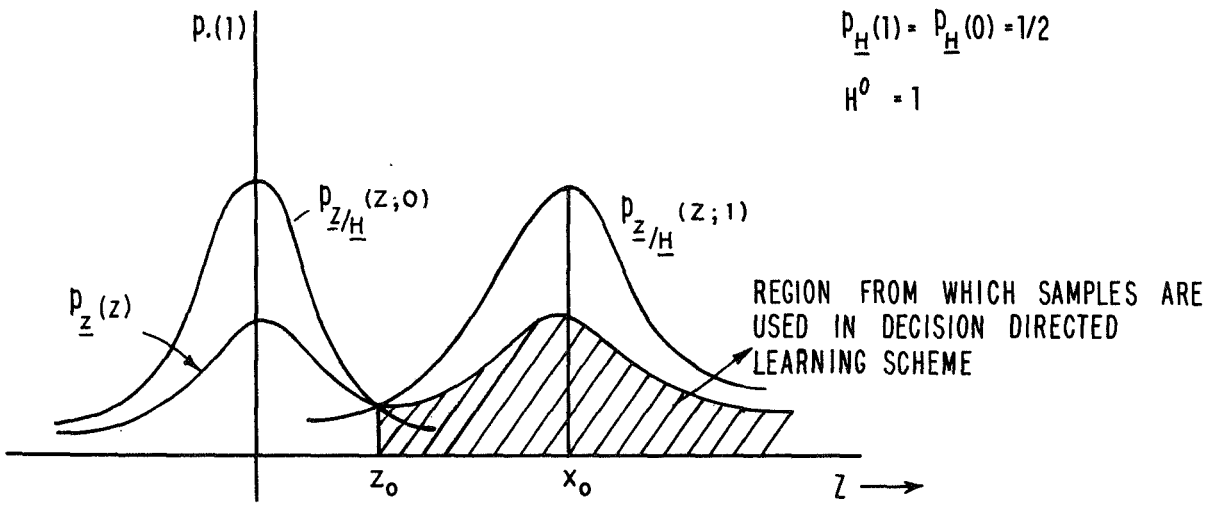


FIG. 1.2 DECISION DIRECTED "LEARNING" REGION

shall be using the samples having the probability density function shown as the shaded area in Figure 2. To learn correctly we should be using the samples having a density function $p_{\underline{z}/\underline{H}}(z; H^o)$. For this example we would expect the Decision Directed Learning scheme to converge to a value of x which is larger than the correct value.

Scudder [9] has analysed the convergence properties of this learning scheme. He points out that there is a finite probability of error in the limit for this learning scheme and hence it does not converge in the sense of (1.20). Patrick [10] points out that the Decision Directed Learning scheme can be used in practice only under high signal to noise ratio problems.*

Therefore we note that the labelling method used in the Decision Directed Learning scheme does not satisfy the second requirement for the labelling process as required in Section I. 5.

I. 6. The Relation of Labelling to the Bayesian Learning Scheme

Let us consider the unsupervised learning problem (Problem B) with a discrete $\mathcal{H}$ space. Let $\mathcal{H}$ space have n points. If we observe a sequence $\mathfrak{z}_k$ consisting of k observations, there are n^k possible sequences for the labels.

In the Bayesian learning scheme we start from the prior $p_{\underline{x}}(x)$ and observing $\underline{z}_1 = z_1$ we compute $p_{\underline{x}/\underline{z}_1}(x; z_1)$ using Eqs. (1.16), (1.17) and (1.18). As $\mathcal{H}$ is discrete the integral in (1.18) can be replaced by a summation and we write (1.16) as

---

*For more about this learning scheme see Appendix A.

$$p_{\underline{x}/\underline{z}}(x; z_1) = \frac{\sum_{i=1}^{n} p_{\underline{z}_1/\underline{H}_1, \underline{x}}(z_1; H_1^i, x) p_{\underline{H}_1}(H_1^i)}{\sum_{j=1}^{n} p_{\underline{z}_1/\underline{H}_1}(z_1; H_1^j) p_{\underline{H}_1}(H_1^j)} \; p_{\underline{x}}(x)$$

$$= \sum_{i=1}^{n} \frac{p_{\underline{z}_1/\underline{H}_1, \underline{x}}(z_1; H_1^i, x) p_{\underline{H}_1}(H_1^i) p_{\underline{x}}(x)}{p_{\underline{z}_1/\underline{H}_1}(z_1; H_1^i) p_{\underline{H}_1}(H_1^i)}$$

$$\cdot \; \frac{p_{\underline{z}_1/\underline{H}_1}(z_1; H_1^i) p_{\underline{H}_1}(H_1^i)}{\sum_{j=1}^{n} p_{\underline{z}_1/\underline{H}_1}(z_1; H_1^j) p_{\underline{H}_1}(H_1^j)}$$

$$= \sum_{i=1}^{n} p_{\underline{x}/\underline{z}_1, \underline{H}_1}(x; z_1, H_1^i) p_{\underline{H}_1/\underline{z}_1}(H_1^i; z_1) \qquad (1.32)$$

The Bayesian learning scheme computes the posterior density function $p_{\underline{x}/\underline{z}_1, \underline{H}_1}(x; z_1, H_1^i)$ along all possible (n) classifications H_1^i of z_1. The posterior density $p_{\underline{x}/\underline{z}_1}(x; z_1)$ is then computed by weighing $p_{\underline{x}/\underline{z}_1, \underline{H}_1}(x; z_1, H_1^i)$ with $p_{\underline{H}_1/\underline{z}_1}(H_1^i; z_1)$. Here $p_{\underline{H}_1/\underline{z}_1}$ is the probability that the classification of z_1 was H_1^i given the observed value of z_1. The same process goes on at every stage. Hence for k observations the Bayesian learning scheme considers n^k possible sequences of labels or classifications and computes the posterior density function by averaging over all the n^k possibilities. The computation of the posterior density function along any sequence of labels is like "learning with a teacher". The number of such sequences is n^k and thus grows exponentially with k.*

*Note that this unsupervised learning scheme has been discussed in general earlier on page 1-15. Here we are considering it for discrete $\mathcal{H}$ space.

If we are given the correct classification of all the observed samples, out of n^k sequences we have to follow only one sequence (the correct sequence) of labels. In any other labelling process we decide one sequence of labels according to our labelling process.

In general, the methods of labelling are limited only by our imagination in finding a method of selecting a classification H_k for the observed z_k. But does there exist a labelling method which can be implemented and which assures the convergence of the resulting learning scheme to the correct value?

In this work we have attempted to answer this question in the affirmative.

I. 7. Summary

In this chapter we have reviewed two learning problems. Problem A is a supervised learning problem while Problem B is an unsupervised learning problem. The Bayesian learning scheme can be used for both of these problems. The solution of the supervised learning problem using Bayesian learning is reasonably simple while the unsupervised Bayesian learning is enormously complex. Through a simple example we have seen how the complexity of the unsupervised Bayesian learning scheme increases exponentially.

Labelling the observed samples in an unsupervised learning problem can lead to a feasible solution. The only labelling method presented in the literature uses the solution of a decision problem as the label and hence results in a 'Decision Directed Learning Scheme'. This learning scheme has some undesirable convergence properties.

In the Bayesian learning scheme for the unsupervised learning problem all possible sequences of the classes are considered. The final posterior density is computed by weighing the posterior densities along all such sequences with the probability of their occurance. By labelling we select one of all such sequences. A question is raised regarding the existence of other possible labelling schemes which assure the convergence.

Appendix A - More about Decision Directed Learning

The term decision directed learning scheme applies to a labelled learning scheme in which a label ℓ_k is generated for the observation z_k by a decision process. This decision process is used to classify z_k to some class H_k and then H_k is treated as the label ℓ_k. The observation z_k and the label ℓ_k are then used in estimation. Therefore to formulate a decision directed learning scheme a decision procedure and an estimation procedure are required.

When formulated in a Bayesian framework* a Bayesian minimum probability of error decision procedure is used to get the label. The observation and the label are then used in Bayesian estimation. Scudder [9] considered a two class problem with two gaussian conditional densities and assumed the mean of one of the two densities as unknown. He constructed a decision directed estimator for this problem in a Bayesian framework and found that asymptotically this estimate does not converge to the correct value.

Patrick, Costello [10], [12], [13], [14], and Monds [11] have considered decision directed learning in other parametric and nonparametric frameworks and have studied general properties of the decision directed estimators. They have indicated the following as some properties of a decision directed estimator:

---

*The decision directed scheme is formulated in this framework in Section I. 5. 1.

1. A decision directed estimator does not converge to a unique value, in general, for a multi class problem. This is because of the presence of various trap states.*
2. This estimator, under very general conditions, converges to a unique value for a two class problem. This unique value is not the correct value, and it depends on the decision procedure used.
3. The decision directed estimator uses samples from a finite or semi infinite region of $\mathcal{Z}$ space (determined by the decision boundaries) in learning about a class. If the conditional densities $(p_{\underline{z}/\underline{H}}(z; H))$ overlap, such regions do not contain the complete conditional density function and hence the estimator has some asymptotic error. As the overlap of the density functions decreases this error becomes small and goes to zero for non-overlapping density functions.
4. The performance of the estimator is very seriously affected by the starting values (a priori information).
5. The main advantage of a decision directed estimator is that it is implementable and gives good cost effective performance under high signal to noise ratio conditions [11].

---

*A trap state is defined as a state at which no further updating is possible for the estimate [14]. This is a point in $\mathcal{X}$ space.

References for Chapter I

1. Wald, A., 'Sequential Analysis,' New York, Wiley, 1947.

2. Ho, Y. C., and Agrawala, A. K., "On Pattern Classification Algorithms -- Introduction and Survey," Proceedings of IEEE, Vol. 56, No. 12, pp. 2101-2114, Dec. 1968.

3. Braverman, D. J., "Machine Learning and Automatic Pattern Recognition," Stanford University, Stanford, California, Tech. Rept. 2003-1, Feb. 1961.

4. Fralick, S. C., "The Synthesis of Machines Which Learn Without a Teacher," Stanford University, Stanford, Calif., Tech. Rept. 6103-8, April 1964.

5. Bharucha, B. H., "A Posterior Distributions and Detection Theory," Information and Control, Vol. 14, pp. 98-132, Jan. 1969.

6. Spragins, J. D., "Reproducing Distributions for Machine Learning," Stanford University, Stanford, Calif., Tech. Rept. 6103-7, Nov. 1963.

7. Raiffa, H., and Schlaifer, R., 'Applied Statistical Decision Theory,' Harvard Business School, Boston, 1961.

8. Bryson, A. E., and Ho, Y. C., 'Applied Optimal Control: Optimization, Estimation and Control,' Waltham, Mass., Blaisdell, 1969.

9. Scudder, H. J., "Probability of Error of Some Adaptive Pattern Recognition Machines," IEEE Trans. on Info. Th., Vol. IT-11, No. 3, July 1965.

10. Patrick, E. A., and Costello, J. P., "Unsupervised estimation and processing of unknown signals," Purdue University Rept. TR-EE 69-18, June 1969.

11. Patrick, E. A., Costello, J. P., and Monds, F. C., "Decision Directed Estimation of a Two Class Decision Boundary," IEEE Trans. on Computers, Vol. C-19, No. 3, pp. 197-205, March 1970.

12. Patrick, E. A., and Costello, J. P., "Asymptotic Probability of Error Using Two Decision Directed Estimators for Two Unknown Mean Vectors," IEEE Trans. on Information Theory, Vol. IT-14, No. 1, pp. 160-162, January 1968.

13. Patrick, E. A., and Costello, J. P., "Bayes Related Solution to Unsupervised Estimation," 1969 Proceedings of National Electronics Conference, pp. 308-310.

14. Patrick, E. A., and Costello, J. P., "Stochastic Approximation and a Class of Decision Directed Estimation Algorithms," (Advance copy of a paper to be published).

15. Agrawala, A. K., and Prabhu, K. P. S., "On an Unsupervised Estimation Problem," Notes on Decision, Control, and Dynamical Systems, Harvard University, (to be published).

CHAPTER II

LEARNING WITH A PROBABILISTIC TEACHER

In Chapter I we have seen how the Bayesian learning solution to the unsupervised learning problem is enormously complex from a computational viewpoint and how we can use labelling to reduce the complexity of a solution. In this chapter we restrict our attention to those unsupervised learning problems for which a feasible labelling method results in an implementable learning scheme. We define such problems as Problem C and proceed to consider a labelling method which uses a random number generator in its implementation. The convergence properties of the resulting learning scheme are established and some examples are presented.

II. 1. A Class of Unsupervised Learning Problems

In Section I. 2. 1 we considered the framework in which we defined two problems, the supervised learning problem (Problem A) and the unsupervised learning problem (Problem B). To define a class of unsupervised learning problems we further assume that

(1) the form of $p_{\underline{y}_k/\underline{x},\underline{\mathcal{Y}}_{k-1}}(y_k; x, \mathcal{Y}_{k-1}) = p_{\underline{z}_k,\underline{H}_k/\underline{x},\underline{\mathcal{Z}}_{k-1},\underline{\mathcal{H}}_{k-1}}(z_k, H_k; x, \mathcal{Z}_{k-1}, \mathcal{H}_{k-1})$ is such that a fixed dimensional sufficient statistic exists for the density function $p_{\underline{x}/\underline{\mathcal{Y}}_k}(x; \mathcal{Y}_k)$.

In this framework we define Problem C as

Problem C - Given a prior density function $p_{\underline{x}}(x)$ and a sequence of observations $\mathcal{Z}_k$ satisfying (1), compute a sequence of posterior density functions such that it converges to the

correct value as the number of observations increases; i.e.

$$\lim_{k\to\infty} p_{\underline{x}/\mathcal{Z}_k}(x;\mathcal{Z}_k:\mathcal{L}_k) = \delta(x - x_o) \qquad \text{w.p. } 1 \qquad (2.1)$$

where $p_{\underline{x}/\mathcal{Z}_k}(x;\mathcal{Z}_k:\mathcal{L}_k)$ is a posterior density function of the sequence.

Problem C defines a class of unsupervised learning problems. The formulation of Problem C is aimed towards a solution using labelling. The updating process of such a solution requires solving a supervised learning problem. Due to the assumption (1) above, this supervised learning problem entertains reproducing densities and hence can be solved easily using the techniques discussed in the Chapter I. Therefore the updating process becomes simple.

To solve Problem C we require a labelling method which assures the convergence in the sense of (2.1). In Section I.5.1 we noted that the decision directed learning scheme does not converge. As a result we cannot accept the decision directed learning scheme as a solution to Problem C.

II.2. Learning with a Probabilistic Teacher

In Section I.6 we noted that in the Bayesian solution of an unsupervised learning problem all possible sequences of labels for the observed sequence $\mathcal{Z}_k$ are considered. The posterior density function is then computed using Eq. (1.32).

Eq. (1.32) suggests a labelling method. Let us treat the label ℓ_k as a random variable $\underline{\ell}_k$. Let this random variable $\underline{\ell}_k$ have a probability density function

$$p_{\underline{\ell}_k}(\ell_k) = p_{\underline{H}_k/\underline{z}_k, \mathcal{Z}_{k-1}}(\ell_k; z_k, \mathcal{Z}_{k-1})$$

This density function can be computed using Eqs. (1.29) and (1.30) and is the same one used in the decision directed scheme. However, we employ it differently in this case. Knowing the density function $p_{\underline{\ell}_k}(\ell_k)$ we generate the label ℓ_k by drawing it as a random number from this density function. With the availability of a random number generator this task is within the reach of a digital computer.

When we generate the labels this way the average posterior density function at any stage will be the same as the posterior density function for the Bayesian learning scheme. Therefore it is reasonable to hope that this in limit, will lead to a solution of Problem C.

In this labelling method the label is generated probabilistically. Therefore we may call the resulting learning scheme as 'Probabilistically Directed Learning Scheme' or 'Learning with a Probabilistic Teacher' (LPT scheme in short). A learning scheme of this type was first suggested in [1]. Its convergence properties were first established in [2].

Let us consider the LPT scheme in detail. The computations for the observation z_k are carried out in two steps, labelling and updating.

II.2.1. The Probabilistic Labelling

At the kth stage of computations, the density function $p_{\underline{x}/\mathcal{Z}_{k-1}, \underline{\mathcal{L}}_{k-1}}(x; \mathcal{Z}_{k-1}, \mathcal{L}_{k-1})$ is available as the prior density function. When we observe z_k we proceed to compute $p_{\underline{H}_k/\underline{z}_k, \mathcal{Z}_{k-1}, \underline{\mathcal{L}}_{k-1}}(H_k; z_k, \mathcal{Z}_{k-1}, \mathcal{L}_{k-1})$ as follows. We write

$$p_{\underline{H}_k/\underline{z}_k,\underline{\mathcal{Z}}_{k-1},\underline{\mathcal{L}}_{k-1}}(H_k; z_k, \mathcal{Z}_{k-1}, \mathcal{L}_{k-1}) =$$

$$\frac{p_{\underline{z}_k,\underline{H}_k/\underline{\mathcal{Z}}_{k-1},\underline{\mathcal{L}}_{k-1}}(z_k, H_k; \mathcal{Z}_{k-1}, \mathcal{L}_{k-1})}{p_{\underline{z}_k/\underline{\mathcal{Z}}_{k-1},\underline{\mathcal{L}}_{k-1}}(z_k; \mathcal{Z}_{k-1}, \mathcal{L}_{k-1})}$$

Under the assumption (A2) we can express it as

$$p_{\underline{H}_k/\underline{z}_k,\underline{\mathcal{Z}}_{k-1},\underline{\mathcal{L}}_{k-1}}(H_k; z_k, \mathcal{Z}_{k-1}, \mathcal{L}_{k-1}) =$$

$$\frac{p_{\underline{z}_k/\underline{H}_k,\underline{\mathcal{Z}}_{k-1},\underline{\mathcal{L}}_{k-1}}(z_k; H_k, \mathcal{Z}_{k-1}, \mathcal{L}_{k-1})\, p_{\underline{H}_k}(H_k)}{p_{\underline{z}_k/\underline{\mathcal{Z}}_{k-1},\underline{\mathcal{L}}_{k-1}}(z_k; \mathcal{Z}_{k-1}, \mathcal{L}_{k-1})} \tag{2.2}$$

Here

$$p_{\underline{z}_k/\underline{\mathcal{Z}}_{k-1},\underline{\mathcal{L}}_{k-1}}(z_k; \mathcal{Z}_{k-1}, \mathcal{L}_{k-1}) =$$

$$\int_{\mathcal{H}} p_{\underline{z}_k/\underline{H}_k,\underline{\mathcal{Z}}_{k-1},\underline{\mathcal{L}}_{k-1}}(z_k; H_k, \mathcal{Z}_{k-1}, \mathcal{L}_{k-1})\, p_{\underline{H}_k}(H_k)\, dH_k \tag{2.3}$$

and

$$p_{\underline{z}_k/\underline{H}_k,\underline{\mathcal{Z}}_{k-1},\underline{\mathcal{L}}_{k-1}}(z_k; H_k, \mathcal{Z}_{k-1}, \mathcal{L}_{k-1}) =$$

$$\int_{\mathcal{X}} p_{\underline{z}_k/\underline{x},\underline{H}_k,\underline{\mathcal{Z}}_{k-1},\underline{\mathcal{L}}_{k-1}}(z_k; x, H_k, \mathcal{Z}_{k-1}, \mathcal{L}_{k-1})$$

$$\cdot\, p_{\underline{x}/\underline{\mathcal{Z}}_{k-1},\underline{\mathcal{L}}_{k-1}}(x; \mathcal{Z}_{k-1}, \mathcal{L}_{k-1})\, dx$$

Under the assumption (A1) of conditional independence we may write it as

$$p_{\underline{z}_k/\underline{H}_k,\underline{\mathfrak{z}}_{k-1},\underline{\mathcal{L}}_{k-1}}(z_k; H_k, \mathfrak{z}_{k-1}, \mathcal{L}_{k-1}) =$$

$$\int_{\underline{X}} p_{\underline{z}_k/\underline{H}_k,\underline{x}}(z_k; H_k, x) p_{\underline{x}/\underline{\mathfrak{z}}_{k-1},\underline{\mathcal{L}}_{k-1}}(x; \mathfrak{z}_{k-1}, \mathcal{L}_{k-1}) dx \tag{2.4}$$

Knowing $p_{\underline{x}/\underline{\mathfrak{z}}_{k-1},\underline{\mathcal{L}}_{k-1}}(x; \mathfrak{z}_{k-1}, \mathcal{L}_{k-1})$ and z_k we can compute the density function $p_{\underline{H}_k/\underline{z}_k,\underline{\mathfrak{z}}_{k-1},\underline{\mathcal{L}}_{k-1}}(H_k; z_k, \mathfrak{z}_{k-1}, \mathcal{L}_{k-1})$ using the Eqs. (2.2), (2.3) and (2.4).* This computation can be carried out for discrete as well as continuous $\mathcal{H}$ spaces. In generating the label ℓ_k we treat it as a random variable $\underline{\ell}_k$ with a probability density function

$$p_{\underline{\ell}_k/\underline{z}_k,\underline{\mathfrak{z}}_{k-1},\underline{\mathcal{L}}_{k-1}}(\ell_k; z_k, \mathfrak{z}_{k-1}, \mathcal{L}_{k-1}) =$$

$$p_{\underline{H}_k/\underline{z}_k,\underline{\mathfrak{z}}_{k-1},\underline{\mathcal{L}}_{k-1}}(\ell_k; z_k, \mathfrak{z}_{k-1}, \mathcal{L}_{k-1}) \tag{2.5}$$

To generate the label we draw a random number from the probability density function $p_{\underline{\ell}_k/\underline{z}_k,\underline{\mathfrak{z}}_{k-1},\underline{\mathcal{L}}_{k-1}}(\ell_k; z_k, \mathfrak{z}_{k-1}, \mathcal{L}_{k-1})$. A pseudo random number generator can be used for this task.+

II. 2. 2. Updating

The updating of the density function requires computing the posterior density function $p_{\underline{x}/\underline{\mathfrak{z}}_k,\underline{\mathcal{L}}_k}(x; \mathfrak{z}_k, \mathcal{L}_k)$ from the prior $p_{\underline{x}/\underline{\mathfrak{z}}_{k-1},\underline{\mathcal{L}}_{k-1}}(x; \mathfrak{z}_{k-1}, \mathcal{L}_{k-1})$ and (z_k, ℓ_k). Using Bayes rule we can express $p_{\underline{x}/\underline{\mathfrak{z}}_k,\underline{\mathcal{L}}_k}(x; \mathfrak{z}_k, \mathcal{L}_k)$ as

---

*Note that Eqs. (2.2), (2.3) and (2.4) are similar to Eqs. (1.29) and (1.30).

+We shall consider the question of generating the label randomly in detail in Section II. 4.

$$p_{\underline{x}/\underline{\mathcal{Z}}_k,\underline{\mathcal{L}}_k}(x;\mathcal{Z}_k,\mathcal{L}_k) =$$

$$\frac{p_{\underline{z}_k,\underline{\ell}_k/\underline{x},\underline{\mathcal{Z}}_{k-1},\underline{\mathcal{L}}_{k-1}}(z_k,\ell_k;x,\mathcal{Z}_{k-1},\mathcal{L}_{k-1})}{p_{\underline{z}_k,\underline{\ell}_k/\underline{\mathcal{Z}}_{k-1},\underline{\mathcal{L}}_{k-1}}(z_k,\ell_k;\mathcal{Z}_{k-1},\mathcal{L}_{k-1})}$$

$$\cdot\ p_{\underline{x}/\underline{\mathcal{Z}}_{k-1},\underline{\mathcal{L}}_{k-1}}(x;\mathcal{Z}_{k-1},\mathcal{L}_{k-1}) \tag{2.6}$$

where

$$p_{\underline{z}_k,\underline{\ell}_k/\underline{\mathcal{Z}}_{k-1},\underline{\mathcal{L}}_{k-1}}(z_k,\ell_k;\mathcal{Z}_{k-1},\mathcal{L}_{k-1}) =$$

$$\int_{\underline{X}} p_{\underline{z}_k,\underline{\ell}_k/\underline{x},\underline{\mathcal{Z}}_{k-1},\underline{\mathcal{L}}_{k-1}}(z_k,\ell_k;x,\mathcal{Z}_{k-1},\mathcal{L}_{k-1})$$

$$\cdot\ p_{\underline{x}/\underline{\mathcal{Z}}_{k-1},\underline{\mathcal{L}}_{k-1}}(x;\mathcal{Z}_{k-1},\mathcal{L}_{k-1})dx \tag{2.7}$$

The updating process defined by Eqs. (2.6) and (2.7) is very similar to the solution of Problem A defined by Eqs. (1.5) and (1.6). For Problem C as defined in Section II.1 some reproducing densities exist for the posterior density function here. As a result the computation of the posterior density $p_{\underline{x}/\underline{\mathcal{Z}}_k,\underline{\mathcal{L}}_k}(x;\mathcal{Z}_k,\mathcal{L}_k)$ from the prior $p_{\underline{x}/\underline{\mathcal{Z}}_{k-1},\underline{\mathcal{L}}_{k-1}}(x;\mathcal{Z}_{k-1},\mathcal{L}_{k-1})$ is reasonably simple.

We note that the updating process for the LPT scheme is the same as the updating for the decision directed scheme. In fact whatever method we may use to generate the label, if we treat ℓ_k as the correct classification for z_k, the updating process will be the same.

II. 2. 3. The Operation of LPT Scheme

The recursive computations for the LPT scheme proceed in two steps at any stage. At the kth stage we observe $\underline{z}_k = z_k$. The density function $p_{\underline{x}/\underline{z}_{k-1},\underline{\mathcal{L}}_{k-1}}(x; \underline{z}_{k-1}, \mathcal{L}_{k-1})$ is available as the prior density function. As a first step we generate a label ℓ_k according to the method described above in Section II. 2. 1. This label ℓ_k along with the observation z_k is used to update the density function $p_{\underline{x}/\underline{z}_{k-1},\underline{\mathcal{L}}_{k-1}}(x; \underline{z}_{k-1}, \mathcal{L}_{k-1})$ to $p_{\underline{x}/\underline{z}_k,\underline{\mathcal{L}}_k}(x; \underline{z}_k, \mathcal{L}_k)$. And we proceed to observe $\underline{z}_{k+1}$.

To start the recursive computations here we require the knowledge of the prior density function $p_{\underline{x}}(x)$ which is used along with the first observation z_1 in the probabilistic labelling at the first stage.

In II. 2. 2 we noted that the updating process for the LPT scheme and the decision directed learning scheme are the same. For labelling, both these schemes require the density function $p_{\underline{H}_k/\underline{z}_k,\underline{z}_{k-1},\underline{\mathcal{L}}_{k-1}}(H_k; z_k, \underline{z}_{k-1}, \mathcal{L}_{k-1})$. This density function is used differently by the two schemes. In the decision directed learning scheme the label ℓ_k^{dd} is assigned that value of H_k for which $p_{\underline{H}_k/\underline{z}_k,\underline{z}_{k-1},\underline{\mathcal{L}}_{k-1}}$ is maximum. The LPT scheme generates the label randomly with the probability density function $p_{\underline{H}_k/\underline{z}_k,\underline{z}_{k-1},\underline{\mathcal{L}}_{k-1}}$. Therefore for an observation z_k the labels assigned by the two schemes will be the same with the probability $p_{\underline{H}_k/\underline{z}_k,\underline{z}_{k-1},\underline{\mathcal{L}}_{k-1}}(\ell_k^{dd}; z_k, \underline{z}_{k-1}, \mathcal{L}_{k-1})$. Note that $p_{\underline{H}_k/\underline{z}_k,\underline{z}_{k-1},\underline{\mathcal{L}}_{k-1}}(\ell_k^{dd}; z_k, \underline{z}_{k-1}, \mathcal{L}_{k-1})$ is the largest value of the function $p_{\underline{H}_k/\underline{z}_k,\underline{z}_{k-1},\underline{\mathcal{L}}_{k-1}}$. Therefore for k observations $\mathcal{L}_k^{dd}$ (the sequence of

labels assigned by the decision directed scheme) has the largest probability of all the sequences of labels the LPT scheme can generate; this probability being

$$p_{\underline{\mathcal{L}}_k/\underline{\mathcal{Z}}_k}(\mathcal{L}_k^{dd};\mathcal{Z}_k) = \prod_{i=1}^{k} p_{\underline{H}_i/\underline{z}_i,\underline{\mathcal{Z}}_{i-1},\underline{\mathcal{L}}_{i-1}}(\ell_i^{dd};z_i,\mathcal{Z}_{i-1},\mathcal{L}_{i-1}) \quad .$$

As each of the multiplying factors on the right hand side of this expression is less than one, the probability $p_{\underline{\mathcal{L}}_k/\underline{\mathcal{Z}}_k}(\mathcal{L}_k^{dd};\mathcal{Z}_k)$ decreases as k increases. For small k this probability may be significant and as this is the probability with which the sequence of labels generated by the LPT scheme and decision directed scheme are the same, the behavior of two schemes may not be significantly different for a small number of observations. But as k increases this probability becomes very small. As a result the convergence properties of the two schemes may be entirely different. In Section I. 5. 1 we noted that the decision directed learning scheme does not converge. Let us examine the convergence properties of the LPT scheme next.

II. 3. The Convergence of LPT Scheme

The way we have defined Problem C in Section II. 1 we require that its solution be in the form of a sequence of posterior density functions which converges in the sense of (2. 1) to a delta function at the correct value. As was pointed out in I. 3. 4 the Bayesian learning scheme offers a solution to this problem. To accept the LPT scheme as a solution we have to establish its convergence.

We require another assumption before we can establish the convergence of the LPT scheme. We assume that

(A4) $$p_{\underline{z}_k/\underline{H}_k,\underline{x},\underline{\mathcal{Z}}_{k-1},\underline{\mathcal{L}}_{k-1}}(z_k; H_k, x, \mathcal{Z}_{k-1}, \mathcal{L}_{k-1}) \neq 0$$

$$\text{for any } x \in \mathcal{X}, \quad z_k \in \mathcal{Z}, \quad H_k \in \mathcal{H}$$

Under the assumption (A1) of conditional independence we may write it as

$$p_{\underline{z}_k/\underline{H}_k,\underline{x}}(z_k; H_k, x) \neq 0 \qquad \text{for any } x \in \mathcal{X}, \quad z_k \in \mathcal{Z}, \quad H_k \in \mathcal{H}$$

The need for this assumption for the LPT scheme arises in the following way. From Eq. (2.6) we note that $p_{\underline{z}_k/\underline{H}_k,\underline{x}}$ is the only function of x which multiplies the prior density function $p_{\underline{x}/\underline{\mathcal{Z}}_{k-1},\underline{\mathcal{L}}_{k-1}}(x; \mathcal{Z}_{k-1}, \mathcal{L}_{k-1})$ in the updating. If we let it have a zero value for some x then there is no way of changing the value of the posterior density function at that value of x due to the later observations. As the LPT scheme assigns labels probabilistically we cannot let a single observation fix the value of all the subsequent posterior density functions for any value of x. Therefore if the assumption (A4) does not hold for some problem the LPT scheme solution may lead to wrong results.

A list of problems which entertain the reproducing densities has been compiled by Spragins [10] and Raiffa and Schlaiffer [12]. Examining their list we find that the assumption (A4) does not put any severe restrictions. <u>It is satisfied by all the problems with the reproducing</u>

densities. The only exception is a problem in which the conditional density $p_{\underline{z}/\underline{H}}$ is a uniform distribution and the range of such uniform distribution is the unknown. A detailed discussion of this problem is presented in Appendix B.

The convergence of the LPT scheme is a direct consequence of two theorems. These theorems which establish the proof of convergence are formally presented in Appendix A. The assumption (A4) is required in the proof of Theorem I. In Theorem 2 we require the further assumption of the existence of a sequence of functions of the observations converging to the correct value. In other words, from whatever we are observing and the way we are assigning the labels, the possibility of reaching the correct value from $p_{\underline{x}}(x)$ is not ruled out. Further, as shown in the proof of Theorem 2 in the Appendix A, the existence of such a sequence of functions implies a unique solution. This assures that the mixture we are dealing with is identifiable [11].*

---

*Note that we are dealing with the mixture

$$p_{\underline{z}}(z) = \int_{\mathcal{H}} p_{\underline{z}/\underline{H}}(z; H) p_{\underline{H}}(H) dH \qquad \text{(A)}$$

If $\mathcal{H}$ has two points only, H^o and H^1, $p_{\underline{z}}(z)$ takes the form

$$p_{\underline{z}}(z) = p_{\underline{z}/\underline{H}}(z; H^o) p_{\underline{H}}(H^o) + p_{\underline{z}/\underline{H}}(z; H^1) p_{\underline{H}}(H^1)$$

We say that this mixture is identifiable if for any two points z^1 and z^2,

$$p_{\underline{z}}(z^1) = p_{\underline{z}/\underline{H}}(z^1; H^o) p_{\underline{H}}(H^o) + p_{\underline{z}/\underline{H}}(z^1; H^1) p_{\underline{H}}(H^1)$$

and

$$p_z(z^2) = p_{\underline{z}/\underline{H}}(z^2; H^o) p_{\underline{H}}(H^o) + p_{\underline{z}/\underline{H}}(z^2; H^1) p_{\underline{H}}(H^1)$$

are two independent equations in $p_{\underline{H}}(H^o)$ and $p_{\underline{H}}(H^1)$.

In general $p_{\underline{z}}(z)$ is said to be identifiable if the mapping of $p_{\underline{H}}(H)$ on to $p_{\underline{z}}(z)$ as defined by the Eq. (A) above, is one to one [11].

To see heuristically why the LPT scheme converges we note that we are treating the label as a random variable. If we consider the average behavior of the posterior density function with respect to the randomness in the label we get

$$\underset{\mathcal{L}_k}{E}\left[p_{\underline{x}/\underline{\mathcal{Z}}_k,\underline{\mathcal{L}}_k}(x;\mathcal{Z}_k,\mathcal{L}_k)\right] = \int p_{\underline{x}/\underline{\mathcal{Z}}_k,\underline{\mathcal{L}}_k}(x;\mathcal{Z}_k,\mathcal{L}_k) \cdot p_{\underline{\mathcal{L}}_k/\underline{\mathcal{Z}}_k}(\mathcal{L}_k;\mathcal{Z}_k)d\mathcal{L}_k \quad . \tag{2.8}$$

The way we have selected the label, the right hand side is nothing but

$$p_{\underline{x}/\underline{\mathcal{Z}}_k}(x;\mathcal{Z}_k) \quad . \tag{2.9}$$

This is the posterior density function for the Bayesian learning scheme. Hence the LPT scheme follows the Bayesian learning scheme on the average.

The convergence of the posterior density functions $p_{\underline{x}/\underline{\mathcal{Z}}_k}(x;\mathcal{Z}_k)$, involved in the Bayesian learning scheme, has been well established [3], that

$$\lim_{k\to\infty} p_{\underline{x}/\underline{\mathcal{Z}}_k}(x;\mathcal{Z}_k) = \delta(x - x_o) \qquad \text{w.p. } 1 \tag{2.10}$$

All the functions $p_{\underline{x}/\underline{\mathcal{Z}}_k,\underline{\mathcal{L}}_k}$ are positive functions. The expectation operation of Eq. (2.8) is like positive summation. The only way in which Eq. (2.9) and (2.10) can be satisfied together is if

$$\lim_{k\to\infty} p_{\underline{x}/\underline{\mathcal{Z}}_k,\underline{\mathcal{L}}_k}(x;\mathcal{Z}_k,\mathcal{L}_k) = \delta(x - x_o) \qquad \text{w.p. } 1 \tag{2.11}$$

II. 4. The Implementation of the LPT Scheme

Let us consider the practical questions regarding the implementation of the LPT scheme. For this scheme at k^{th} stage

(i) given z_k and $p_{\underline{x}/\mathcal{Z}_{k-1},\underline{\mathcal{L}}_{k-1}}$ we compute $p_{\underline{H}_k/\underline{z}_k,\mathcal{Z}_{k-1},\underline{\mathcal{L}}_{k-1}}$;

(ii) with $p_{\underline{\ell}_k/\underline{z}_k,\mathcal{Z}_{k-1},\underline{\mathcal{L}}_{k-1}}(\ell_k; z_k,\mathcal{Z}_{k-1},\mathcal{L}_{k-1}) = p_{\underline{H}_k/\underline{z}_k,\mathcal{Z}_{k-1},\underline{\mathcal{L}}_{k-1}}(\ell_k; z_k,\mathcal{Z}_{k-1},\mathcal{L}_{k-1})$ (Eq. 2.5) we draw a random number $\underline{\ell}_k$ having this probability density; and

(iii) using z_k and ℓ_k we update the prior density function $p_{\underline{x}/\mathcal{Z}_{k-1},\underline{\mathcal{L}}_{k-1}}(x;\mathcal{Z}_{k-1},\mathcal{L}_{k-1})$ to the posterior density function $p_{\underline{x}/\mathcal{Z}_k,\underline{\mathcal{L}}_k}(x;\mathcal{Z}_k,\mathcal{L}_k)$.

The way we have defined Problem C, the prior and posterior density functions here entertain a sufficient statistic. As a result step (iii) requires recomputing the values of the parameters defining the sufficient statistic. This is a straightforward operation. Step (i) requires computing $p_{\underline{H}_k/\underline{z}_k,\mathcal{Z}_{k-1},\underline{\mathcal{L}}_{k-1}}(H_k; z_k,\mathcal{Z}_{k-1},\mathcal{L}_{k-1})$. We know the functional forms and the parameter values of all the functions involved in this computation (Eqs. 2.2, 2.3 and 2.4). Therefore we can carry out this step also.

We are treating the label ℓ_k as a random variable $\underline{\ell}_k$. In step (i) we compute the probability density function we want $\underline{\ell}_k$ to have. In step (ii) we have to generate a random number having this probability density function. Note that we want a single outcome or observation of the random variable $\underline{\ell}_k$ as ℓ_k.

To examine various techniques we can use for this, let us consider the problem of generating an outcome of the random variable $\underline{\ell}$ defined in $\mathcal{H}$ space and having a probability density function $p_{\underline{\ell}}(\ell)$.

Algorithms are available [4] for generating Pseudo random number $\underline{\omega}$ on a digital computer. $\underline{\omega}$ has a uniform probability density on $\Omega = [0, 1]$. To generate any other random variable on the computer we can consider ω as a random sample point (from the sample space Ω), and define the random variable $\underline{\ell}$ as

$$\underline{\ell} = \psi(\underline{\omega}) \quad . \tag{2.12}$$

To get an outcome ℓ we observe a value ω of $\underline{\omega}$. If we know the function ψ we can get ℓ using (2.12). Let us see how we can arrive at the function ψ knowing $p_{\underline{\ell}}(\ell)$ for various $\mathcal{H}$ spaces.

(a) Discrete $\mathcal{H}$ space -

When $\mathcal{H}$ space is discrete $p_{\underline{\ell}}(\ell)$ is a collection of delta functions;

$$p_{\underline{\ell}}(\ell) = \sum_i P(\ell^i)\delta(\ell - \ell^i) \tag{2.13}$$

Here $P(\ell^i)$ gives the probability of occurance of ℓ^i, a point in $\mathcal{H}$ space. Note that

$$\sum_i P(\ell^i) = 1 \tag{2.14}$$

We can easily divide Ω into n parts (where $\mathcal{H}$ contains n points) such that the length of the ith part is $P(\ell^i)$. Now we select ℓ^k if the randomly observed value of $\underline{\omega}$ lies in the region of $P(\ell^k)$. The function ψ now has the form shown in Figure 2.1.

(b) $\mathcal{H}$ is the real line -

Let us define the distribution function $P_{\underline{\ell}}(\ell)$ for the density function $p_{\underline{\ell}}(\ell)$ as

$$P_{\underline{\ell}}(\ell) = \int_{-\infty}^{\ell} p_{\underline{\ell}}(\ell^1)d\ell^1 \tag{2.15}$$

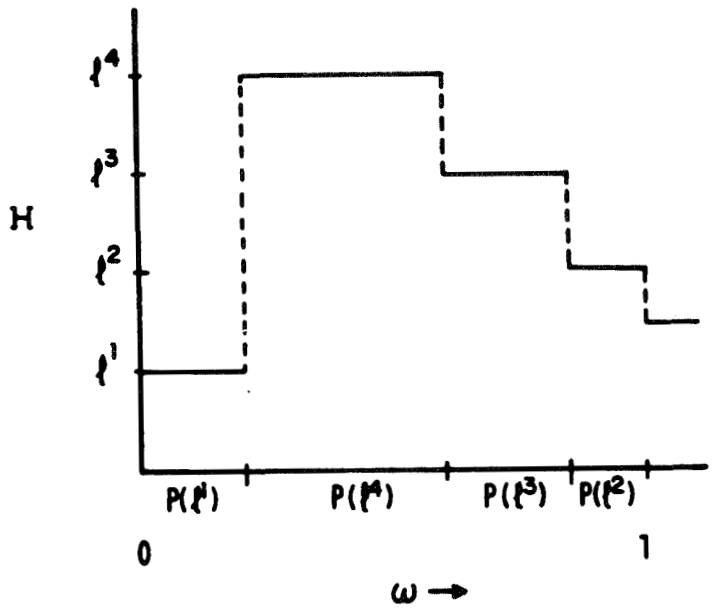


FIG. 2.1 A ψ FUNCTION FOR DISCRETE $\mathcal{H}$

We note that $P_{\underline{\ell}}(\ell)$ is a monotonically increasing function with

$$P_{\underline{\ell}}(-\infty) = 0 \qquad P_{\underline{\ell}}(\infty) = 1$$

We can generate a value of $\underline{\ell}$ by solving

$$P_{\underline{\ell}}(\ell) = \omega \tag{2.16}$$

for any random outcome ω or $\underline{\omega}$. (2.16) is an implicit equation. Reference [9] has presented various techniques for solving such equations. For our purposes we assume that a solution can be found to this equation.

(c) $\mathcal{H}$ is an m dimensional vector space[+] -

If the dimensionality of $\mathcal{H}$ is m we may make an m dimensional sample space as $\Omega_1 \times \Omega_2 \times \Omega_3 \times \ldots \times \Omega_m$ where

$$\Omega_i = [0, 1] \qquad \text{for all } i$$

The random observation is made in this space now as an m dimensional vector. To generate $\underline{\ell}$ we use a function to map this sample space on $\mathcal{H}$*.

So, we see that the LPT scheme leads to a solution of Problem C which can be implemented in practice using a random number generator.

Let us consider some examples next.

II. 5. Examples

Let us consider two examples and examine the behavior of their LPT solution. As the first example we consider the problem of

---

[+] In pattern classification problems the $\mathcal{H}$ space is discrete in general. Therefore this case rarely occurs in pattern classification.

[*] This may be an involved problem but can be solved.

Section I. 4. 1 in which $\mathcal{H}$ is $[H^o, H^1]$. The two conditional densities $p_{\underline{z}/\underline{H}}(z; H^1)$ and $p_{\underline{z}/\underline{H}}(z; H^o)$ are gaussian. The mean of the first conditional density $p_{\underline{z}/\underline{H}}(z; H^1)$ is considered unknown. As the second example we consider the same problem but assume that the variance of one conditional density is unknown.

II. 5. 1. Example C-1

As the first example of the LPT scheme let us consider the problem of Section I. 4. 1 and define Problem C-1 as follows:

Problem C-1

Let

$$p_{\underline{H}_k}(H^1) = p$$

$$p_{\underline{H}_k}(H^o) = 1 - p$$

$$p_{\underline{z}_k/\underline{H}_k}(z_k; H^1) \sim N(x_o, R)$$

$$p_{\underline{z}_k/\underline{H}_k}(z_k; H^o) \sim N(0, R) \tag{2.17}$$

for all k.* (In other words the two conditional densities are gaussian.) We treat the mean of the conditional density function $p_{\underline{z}_k/\underline{H}_k}(z_k; H^1)$ as the unknown parameter x. Observing a sequence $\mathfrak{z}_k$ we have to compute the sequence of the posterior density functions using the LPT scheme.

---

*This structure can also be defined as follows. Let

$$\underline{z}_k = \underline{H}_k x + v_k$$

where

$$\underline{H}_k = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1 - p \end{cases}$$

v_k is a white noise sequence $p_{\underline{v}_k}(v_k) = p_{\underline{v}}(v) \sim N(0, R)$

$\underline{H}_k$ and $\underline{v}_k$ are independent.

Raiffa and Schlaifer [12] have shown that this problem entertains reproducing densities. The reproducing density has the gaussian form,

$$p_{\underline{x}/\underline{z}_k,\underline{\mathcal{L}}_k}(x;\underline{z}_k,\mathcal{L}_k) \sim N(\bar{x}_k, P_k)$$

The mean $\bar{x}_k$ and the variance P_k constitute the fixed dimensional sufficient statistics. The updating, therefore, requires computing the values of these two parameters which can be done as follows,

$$\bar{x}_k = \bar{x}_{k-1} + P_k L_k^T R^{-1}(z_k - L_k \bar{x}_{k-1}) \tag{2.18}$$

$$P_k^{-1} = P_{k-1}^{-1} + L_k^T R^{-1} L_k \tag{2.19}$$

where

$$L_k = 1 \text{ if } \ell_k = H^1 \text{ and } L_k = 0 \text{ if } \ell_k = H^o \quad . \tag{2.20}$$

Eq. (2.18) simply computes the sample mean of all the samples which are labelled as H^1. Here $\bar{x}_k$ is a function of $\underline{z}_k$ and $\mathcal{L}_k$ while P_k is a function of $\mathcal{L}_k$.

The updating for this problem is straightforward. We start with the prior density function $p_{\underline{x}}(x) \sim N(\bar{x}_o, P_o)$ which we assume given. Before we can update, however, we have to generate the label ℓ_k for the observed sample z_k. Let us see how this is done for the LPT scheme.

Labelling - We note that for this problem

$$p_{\underline{z}_k/\underline{x},\underline{H}_k}(z_k; x, \ell_k) = \frac{1}{\sqrt{2\pi R}} e^{-\frac{1}{2R}(z_k - L_k x)^2} \tag{2.21}$$

where L_k is given by (2.20). Therefore using Eq. (2.4)

$$p_{\underline{z}_k/\underline{H}_k,\underline{\mathcal{Z}}_{k-1},\underline{\mathcal{L}}_{k-1}} = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi R}} e^{-\frac{1}{2R}(z_k - L_k x)^2} \cdot \frac{1}{\sqrt{2\pi P_{k-1}}} e^{-\frac{1}{2P_{k-1}}(x - \bar{x}_{k-1})^2} dx$$

$$= \frac{1}{\sqrt{2\pi(R + L_k P_{k-1} L_k^T)}} e^{-\frac{1}{2(R+L_k P_{k-1} L_k^T)}(z_k - L_k \bar{x}_{k-1})^2} \tag{2.22}$$

And

$$p_{\underline{\ell}_k/\underline{z}_k,\underline{\mathcal{Z}}_{k-1},\underline{\mathcal{L}}_{k-1}}(H^1; z_k, \mathcal{Z}_{k-1}, \mathcal{L}_{k-1}) =$$

$$p_{\underline{H}_k/\underline{z}_k,\underline{\mathcal{Z}}_{k-1},\underline{\mathcal{L}}_{k-1}}(H^1; z_k, \mathcal{Z}_{k-1}, \mathcal{L}_{k-1})$$

$$= \frac{\frac{p}{\sqrt{2\pi(R + P_{k-1})}} e^{-\frac{1}{2(R+P_{k-1})}(z_k - \bar{x}_{k-1})^2}}{\frac{p}{\sqrt{2\pi(R + P_{k-1})}} e^{-\frac{(z_k - \bar{x}_{k-1})^2}{2(R+P_{k-1})}} + \frac{(1-p)}{\sqrt{2\pi R}} e^{-\frac{1}{2R}(z_k)^2}}$$

$$= \alpha_k \quad \text{say} \tag{2.23}$$

As we know everything on the right hand side of Eq. (2.23) we can easily compute the value of α_k for any observed z_k. Note that

$$p_{\underline{\ell}_k/\underline{z}_k,\underline{\mathcal{Z}}_{k-1},\underline{\mathcal{L}}_{k-1}}(H^0; z_k, \mathcal{Z}_{k-1}, \mathcal{L}_{k-1}) = 1 - \alpha_k$$

To generate the value for ℓ_k we draw a random number ω from a uniform distribution on $[0, 1]$. If

$$\omega \leqslant \alpha_k \quad , \qquad \ell_k = H^o$$

and if

$$\omega > \alpha_k \quad , \qquad \ell_k = 0 \quad . \tag{2.24}$$

The value of ℓ_k so generated is used as the label in this scheme.* This label is used with z_k in the updating.

The implementation of the LPT solution of this problem is straightforward. To examine the behavior of the solution we considered the following numerical values for the parameters:

$$R = 5$$

$$p = 1/2$$

$$p_{\underline{x}} = N(0, 20)$$

$$x_o = -4$$

The 'learning with a teacher', decision directed and the LPT schemes were simulated on a general purpose digital computer. The sequences of $\bar{x}_k$, the mean of the kth posterior density function, are shown in Figure 2.2 for two typical sequences of 500 observations each. The sequence of $\bar{x}_k$ is plotted for each of the three schemes.

Examining Figure 2.2 we note that the $\bar{x}_k$ sequence for decision directed scheme converges around the value -4.6 while the 'learning with a teacher' and the LPT schemes converge to the value -4.0 and show a very similar behavior.

---

*In the decision directed scheme we choose the label as

$$\ell_k = H^o \quad \text{if} \quad \alpha_k > (1 - \alpha_k) \quad \text{or} \quad \alpha_k \geqslant 0.5$$

$$\ell_k = 0 \quad \text{if} \quad \alpha_k < 0.5 \quad .$$

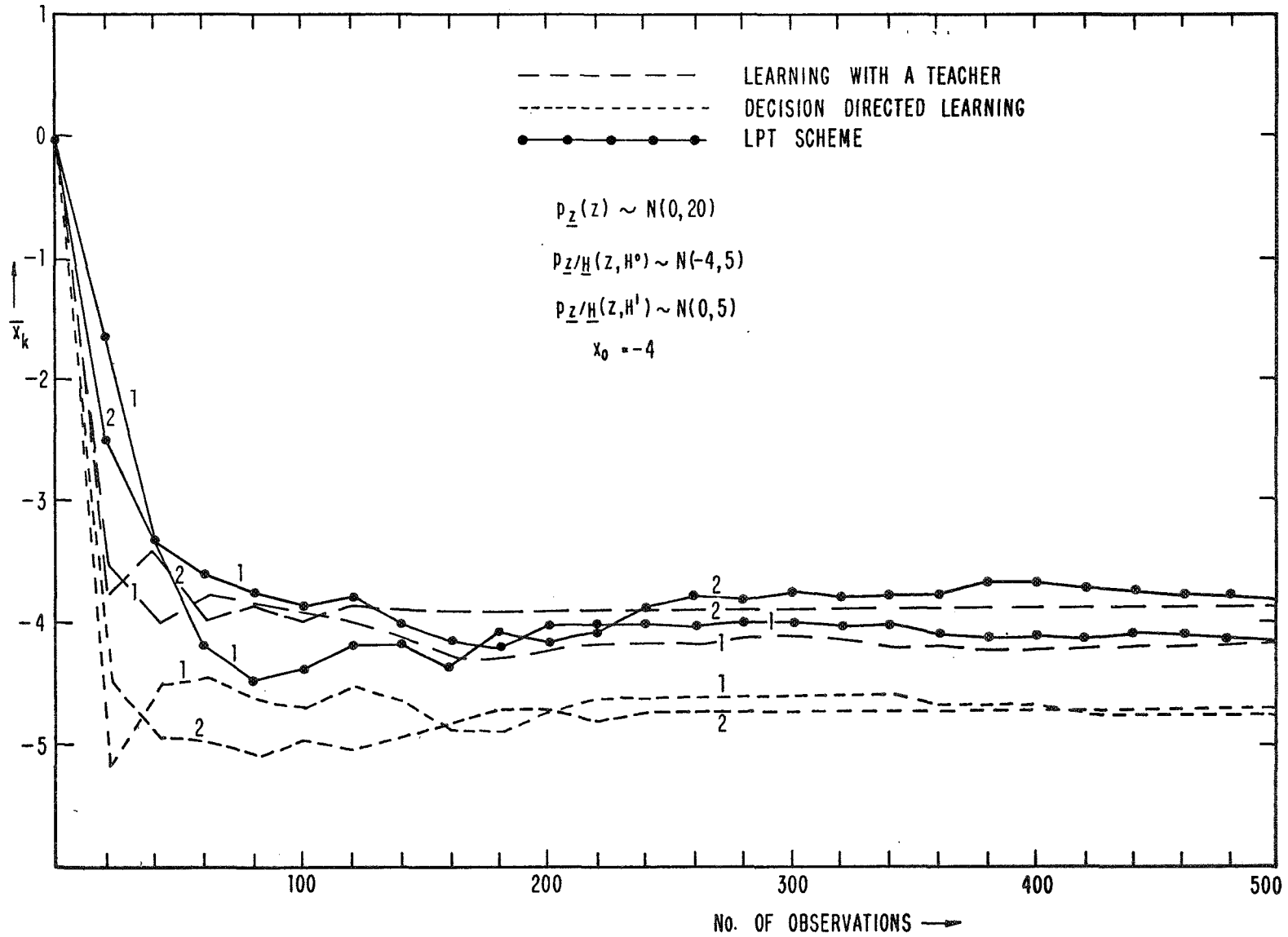


FIG. 2.2 TWO TYPICAL LEARNING SEQUENCES OF $\bar{x}_k$ FOR EXAMPLE C-1

The Bayesian solution ('without a teacher') to the problem considered in this example is infeasible due to the computational complexities discussed in Chapter I. Therefore we cannot compare the performance of the LPT scheme with it. We may, however, compare the performance of the LPT scheme with the 'learning with a teacher' scheme via simulations. For this we would like to get some idea about the variance of $\bar{x}_k$ for these schemes.

To get the sample variance of $\bar{x}_k$ for these schemes we repeated the simulation 60 times starting with $p_{\underline{x}}(x) \sim N(-4, 20)$. The sample variance was computed from these runs* and is plotted in Figure 2.3. In this figure we have omitted the variance curve for the decision directed scheme. As the decision directed scheme converges to an incorrect value, the sample variance curve for this scheme is not meaningful.

Examining Figure 2.3 we note that the variance curves for the LPT scheme and the 'learning with a teacher' scheme have a similar shape. The variance for the LPT scheme is larger than the variance for the 'learning with a teacher' scheme for the same number of observations. We expect this because the 'learning with a teacher' scheme makes use of the knowledge of the correct classifications of observed samples.

Further, we note that the variance of $\bar{x}_k$ for the LPT scheme is approximately twice the variance for the 'learning with a teacher' scheme. We will have more to say on this in Section II.6.

---

*Actually we computed the variance from 60 as well as 80 runs and found them to be very similar.

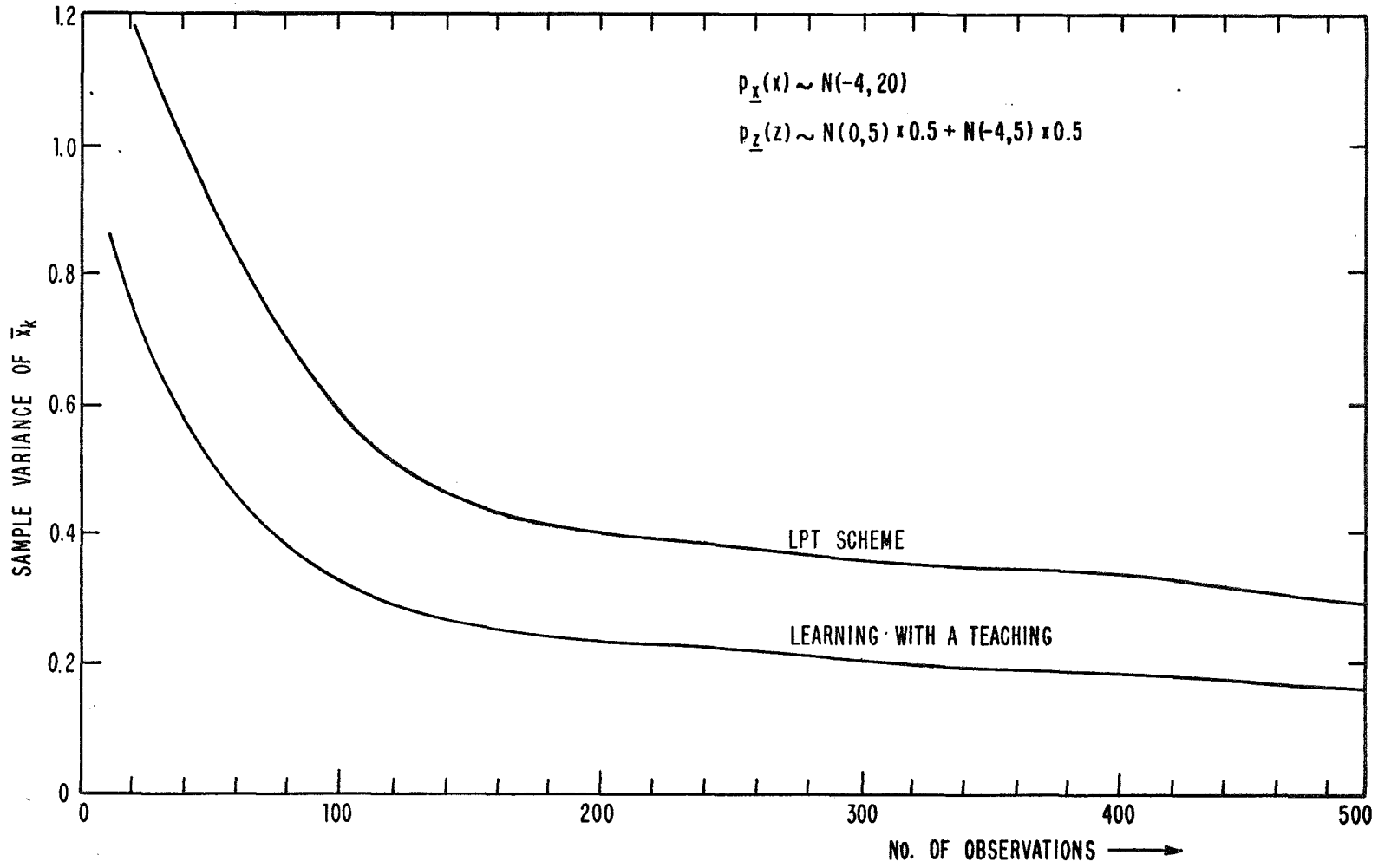


FIG. 2.3 SAMPLE VARIANCE OF $\bar{x}_k$

II. 5. 2. Example C-2

In the first example (C-1) of the LPT scheme we considered a two class case (i. e. $\mathcal{H} = [H^o, H^1]$) in which the two conditional densities were normal and the mean of one of them was unknown. Here let us assume the same structure but consider one of the variances unknown and define Problem C-2 as follows.

Problem C-2

Let

$$p_{\underline{H}_k}(H^1) = p$$

$$p_{\underline{H}_k}(H^o) = 1 - p$$

$$p_{\underline{z}_k/\underline{H}_k}(z_k; H^1) \sim N(\mu, v^o)$$

$$p_{\underline{z}_k/\underline{H}_k}(z_k; H^o) \sim N(0, R) \tag{2.24}$$

for all k. We treat the variance of the conditional density function $p_{\underline{z}_k/\underline{H}_k}(z_k; H^1)$ as the unknown parameter, and are required to formulate the LPT solution for this unknown parameter, observing the sequence $\mathcal{Z}_k$.

If we define the unknown parameter x as 1/variance we can make use of the reproducing densities as described by Raiffa and Schlaiffer [12]. When we define x this way we can write

$$p_{\underline{z}_k/\underline{H}_k, \underline{x}}(\underline{z}_k; H^1, x) = \sqrt{\frac{x}{2\pi}}\, e^{-\frac{1}{2}x(z_k-\mu)^2} \tag{2.25}$$

The Gamma-2 density function is a reproducing density function for

this problem. This density function has the following form:*

$$p_{\underline{x}/\mathfrak{z}_k,\mathcal{L}_k}(x;\mathfrak{z}_k,\mathcal{L}_k) = \frac{\frac{1}{2}\nu_k v_k}{(\frac{1}{2}\nu_k - 1)!}\left(\frac{1}{2}x\nu_k v_k\right)^{\frac{1}{2}\nu_k - 1} e^{-\frac{1}{2}x\nu_k v_k} \tag{2.26}$$

ν_k and v_k are the two parameters of this density function and form the fixed dimensional sufficient statistics. Here ν_k is a function of $\mathcal{L}_k$ alone while v_k is a function of $\mathfrak{z}_k$ and $\mathcal{L}_k$ in the following way:

$$\nu_k = \begin{cases} \nu_{k-1} + 1 & \text{if } \ell_k = H^1 \\ \nu_{k-1} & \text{if } \ell_k = H^o \end{cases} \tag{2.27}$$

$$v_k = \begin{cases} \frac{1}{\nu_k}[\nu_{k-1}v_{k-1} + (z_k - \mu)^2] & \text{if } \ell_k = H^1 \\ v_{k-1} & \text{if } \ell_k = H^o \end{cases} \tag{2.28}$$

Note that by Eq. (2.28) v_k computes the sample variance of all the observations which have a label H^1. Eqs. (2.27) and (2.28) define the updating process for this problem. This updating process starts from a prior density function which is a Gamma-2 distribution with the form,

$$p_{\underline{x}}(x) = \frac{\frac{1}{2}\nu_o v_o}{(\frac{1}{2}\nu_o - 1)!}\left(\frac{1}{2}x\nu_o v_o\right)^{\frac{1}{2}\nu_o - 1} e^{-\frac{1}{2}x\nu_o v_o} \qquad \begin{matrix}\nu_o > 0 \\ v_o > 0\end{matrix}$$

We assume that ν_o and v_o are given to specify the prior density function.

---

*This density function is defined with

$$x > 0\ , \quad \nu_k > 0 \text{ and } v_k > 0 \quad .$$

The first and second moments of this density function are $\frac{1}{v_k}$ and $\frac{1}{\frac{1}{2}\nu_k v_k^2}$ respectively. Figure 2.4 shows a family of curves for this distribution.

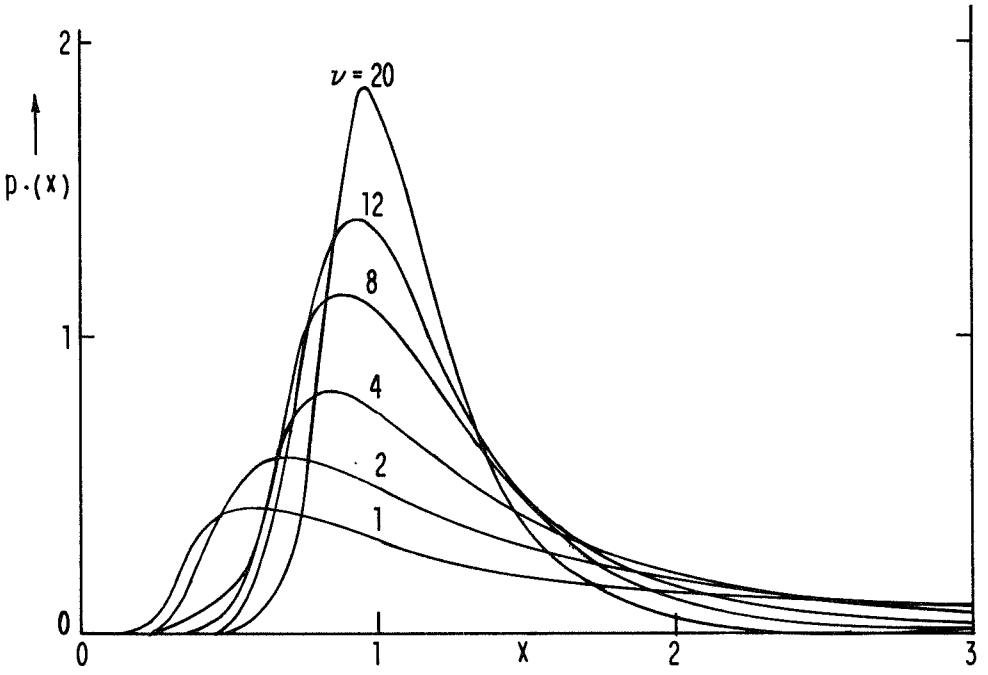


FIG. 2.4 GAMMA-2 DENSITIES, v = 1

$$p_{\underline{x}}(x) = \frac{\frac{1}{2}\nu v}{(\frac{1}{2}\nu - 1)!} (\tfrac{1}{2} x \nu v)^{\frac{1}{2}\nu - 1} \; e^{-\frac{1}{2} x \nu v}$$

We require the label ℓ_k in the updating process. To generate a label using the LPT scheme we have to compute $p_{\underline{z}_k/\underline{H}_k, \mathfrak{z}_{k-1}, \underline{\mathcal{L}}_{k-1}}(\underline{z}_k; H^1, \mathfrak{z}_{k-1}, \mathcal{L}_{k-1})$. Using Eq. (2.4) we write

$$p_{\underline{z}_k/\underline{H}_k, \mathfrak{z}_{k-1}, \underline{\mathcal{L}}_{k-1}}(z_k; H^1, \mathfrak{z}_{k-1}, \mathcal{L}_{k-1}) =$$

$$\int_0^\infty \sqrt{\frac{x}{2\pi}}\, e^{-\frac{1}{2}x(z_k-\mu)^2} \frac{\frac{1}{2}\nu_{k-1}v_{k-1}}{(\frac{1}{2}\nu_{k-1}-1)!}\left(\frac{1}{2}x\nu_{k-1}v_{k-1}\right)^{\frac{1}{2}\nu_{k-1}-1}$$

$$\cdot\, e^{-\frac{1}{2}x\nu_{k-1}v_{k-1}}\, dx$$

$$= \frac{\frac{1}{2}\nu_{k-1}v_{k-1}\left(\frac{1}{2}\nu_{k-1}v_{k-1}\right)^{\frac{1}{2}\nu_{k-1}-1}\left(\frac{1}{2}\nu_{k-1}-\frac{1}{2}\right)!}{\sqrt{2\pi}\left(\frac{1}{2}\nu_{k-1}-1\right).\frac{1}{2}\left(\nu_{k-1}v_{k-1}+(z_k-\mu)^2\right)^{\frac{1}{2}\nu_{k-1}+\frac{1}{2}}}$$

$$= \beta_k \quad \text{say}$$

And

$$p_{\underline{\ell}_k/\underline{z}_k, \mathfrak{z}_{k-1}, \underline{\mathcal{L}}_{k-1}}(H^1; z_k, \mathfrak{z}_{k-1}, \mathcal{L}_{k-1}) = \frac{\beta_k p}{\beta_k p + \frac{(1-p)}{\sqrt{2\pi R}}\, e^{-\frac{1}{2R}(z_k)^2}}$$

$$= \alpha_k \quad \text{say} \quad .$$

Therefore

$$p_{\underline{\ell}_k/\underline{z}_k, \mathfrak{z}_{k-1}, \underline{\mathcal{L}}_{k-1}}(H^0; z_k, \mathfrak{z}_{k-1}, \mathcal{L}_{k-1}) = 1 - \alpha_k \quad .$$

To generate the value for ℓ_k we draw a random number ω from a uniform distribution on $[0, 1]$. If

$$\omega \leq a_k \qquad \ell_k = H^1$$

and if

$$\omega > a_k \qquad \ell_k = H^o$$

The value of ℓ_k so generated can be used in the updating process.

This example was simulated on a general purpose digital computer with the following parameter values;

$p = 0.5$

$R = 5$

$\mu = 0$

$\nu_o = 1$

$v_o = 1$

The correct value of x was taken as 0.1 so that the correct value of the variance was 10.0.

Figure 2.5 shows two typical sequences of v_k for the LPT, decision directed and 'learning with a teacher' schemes. The decision directed solution shows a very eratic behavior* while the v_k sequence converges to the correct value, 10, for both the LPT and 'learning with a teacher' schemes. To get a better idea about the behavior of the LPT and 'learning with a teacher' solutions we repeated the simulation 50 times and computed the sample variance of

---

*Among the many trial runs we found that the v_k sequence of the decision directed solution to this problem either remains at a very small value -- around 2.4 -- or goes to a very large value around 12.5. This happens often enough that the two curves shown represent very typical cases of the decision directed solution for this problem.

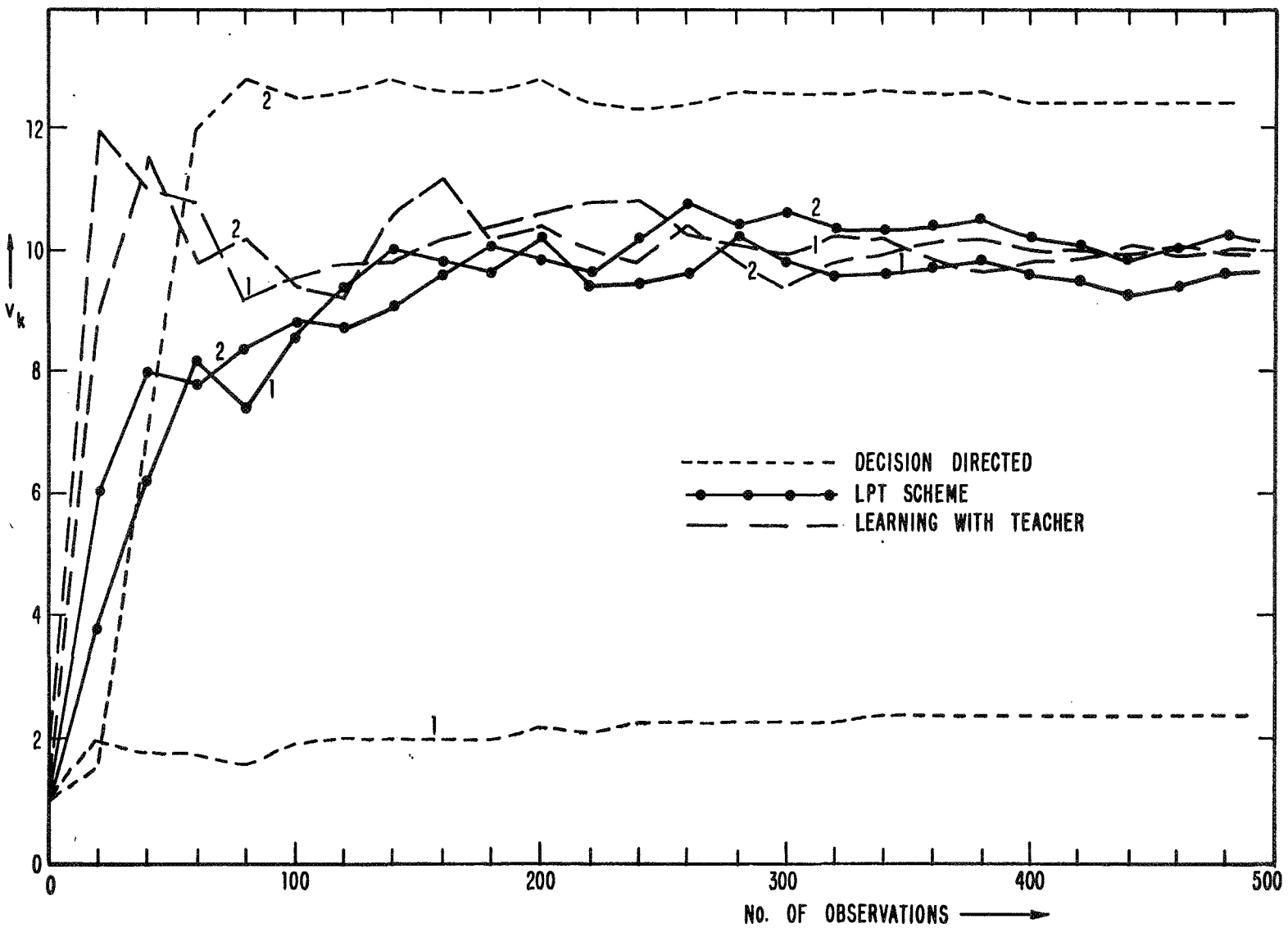


FIG. 2.5 TWO TYPICAL LEARNING SEQUENCES OF v_k FOR PROBLEM C-2

v_k for both the schemes. This sample variance is shown in Figure 2.6.+

Examining the Fig. 2.6 we find that all the comments about Fig. 2.3 in Section II.5.1 are valid here also. Here the variance of v_k for the LPT scheme becomes approximately twice the variance of v_k for the 'learning with a teacher' scheme, as the number of observations increases. We shall try to explain this in the next section.

II.6. The LPT Estimate: Some Properties

The structure of the problems we are considering here is such that x is some unknown parameter of the conditional density function $p_{\underline{z}/\underline{H}}(z; H:x)$. We are given some observations and have to make an estimate of x which is optimal with respect to the cost function J defined by Eq. (1.4). When using the Bayesian estimation philosophy we found that the posterior density function contains all the information required for making the optimal estimate. Therefore we reformulated the supervised and unsupervised learning problems (Problem A and B) of Section I.2.1 in Section I.3 and accepted the posterior density function as a solution. In Section II.1 we restricted our attention to a class of unsupervised learning problems and called this Problem C. We accepted a sequence of posterior density functions which converge in the sense of (2.1) as a solution to Problem C. As a result we get the posterior density function $p_{\underline{x}/\underline{z}_k, \underline{h}_k}(x; z_k, h_k)$ as a solution to Problem A and $p_{\underline{x}/\underline{z}_k, \underline{\mathcal{L}}_k}(x; z_k, \mathcal{L}_k)$ as a solution to Problem C. From the posterior density function we are required to make an estimate of x.

+We calculated the variance for 50 as well as 80 runs and found the two to be very similar.

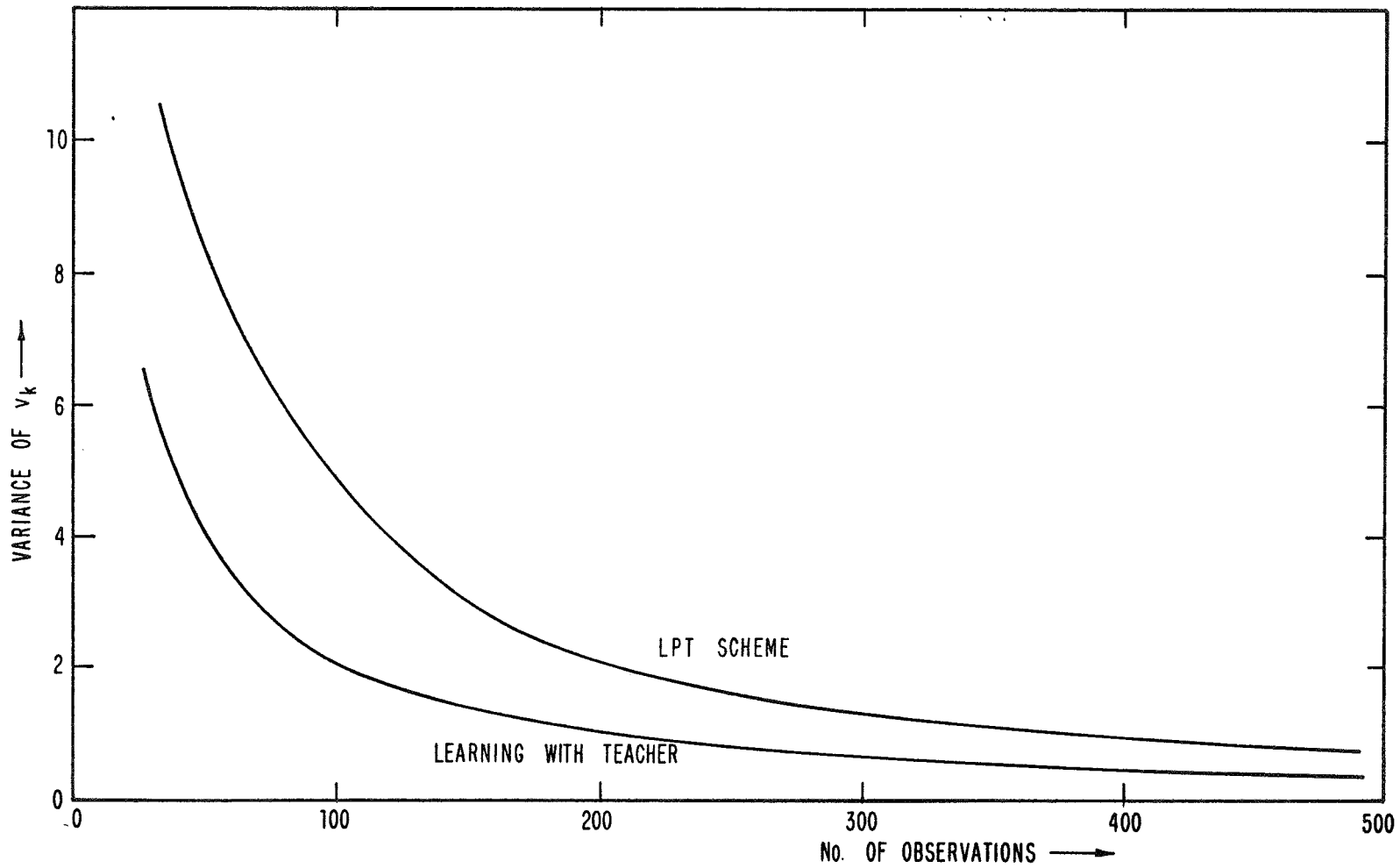


FIG. 2.6 SAMPLE VARIANCE OF v_k

To make an optimal estimate using the posterior density function $p_{\underline{x}/\mathcal{Z}_k, \underline{h}_k}(x; \mathcal{Z}_k, \mathcal{h}_k)$ or $p_{\underline{x}/\mathcal{Z}_k, \underline{\mathcal{L}}_k}(x; \mathcal{Z}_k, \mathcal{L}_k)$ we require the knowledge of the cost function J. In general $\hat{x}_k^A$, the estimate in the Problem A after k observations, is a function of the k observations $(\mathcal{Z}_k, \mathcal{h}_k)$. Therefore we may write

$$\hat{x}_k^A = \psi_k(\mathcal{Z}_k, \mathcal{h}_k) \quad . \tag{2.29}$$

We note that the density functions $p_{\underline{x}/\mathcal{Z}_k, \underline{h}_k}$ and $p_{\underline{x}/\mathcal{Z}_k \underline{\mathcal{L}}_k}$ have the same form. Therefore if we use the same cost function for the Problem C we will get the same function ψ_k for the estimate $\hat{x}_k^C$ and we may write

$$\hat{x}_k^C = \psi_k(\mathcal{Z}_k, \mathcal{L}_k) \tag{2.30}$$

The estimates here are functions of the random observations and therefore are random variables themselves. We would like to get some idea about the variance of these estimates.

From (2.29) and (2.30) we note that $\hat{x}_k^A$ and $\hat{x}_k^C$ have the same functional form. $\hat{x}_k^A$ uses the correct classifications $\mathcal{h}_k$ while $\hat{x}_k^C$ uses the labels generated in the LPT scheme as the classifications. Note that $\underline{H}_k$ is a random variable and given a value of $\underline{z}_k$, $\underline{H}_k$ has the density function $p_{\underline{H}_k/\underline{z}_k}(H_k; z_k : x^o)$. The label $\underline{\ell}_k$ is generated in the LPT scheme from the density function $p_{\underline{H}_k/\underline{z}_k, \mathcal{Z}_{k-1}, \underline{\mathcal{L}}_{k-1}}$ and is independent of $\underline{H}_k$. We have seen that the LPT scheme converges to the correct value in the sense of (2.1). Therefore $\underline{H}_k$ and $\underline{\ell}_k$ have the same probability density function asymptotically. As a result $\hat{x}_k^A$ and $\hat{x}_k^C$ have identical density functions asymptotically.

<u>Lemma</u>: If $\underline{x}_1$ and $\underline{x}_2$ are two independent random variables with the same density function say $p_{\underline{x}}$ then

$$E\{(x_2 - x_1)^2\} = 2 \operatorname{Var} x_1 \tag{2.31}$$

Proof.

$$E\{(x_2 - x_1)^2\} = E\{x_2^2 - 2x_1x_2 + x_1^2\} = E(x_2^2) - 2E(x_2)E(x_1) + E(x_1^2)$$

$$= 2\{E(x_1^2) - (E(x_1))^2\}$$

$$= 2\,\mathrm{Var}\,x_1$$

If we are trying to estimate the random variable $\underline{x}_1$ and we estimate it by the random variable $\underline{x}_2$ then $(\underline{x}_2 - \underline{x}_1)$ is the error. As $\underline{x}_1$ and $\underline{x}_2$ have identical distributions, the mean error is zero. $E[(x_2 - x_1)^2]$ now has the interpretation of the variance of the error. We see that the variance of the error is twice the variance of $\underline{x}_1$.

$\underline{H}_k$ is an unknown random variable for the LPT scheme in which we are using $\underline{\ell}_k$ in its place. As we have seen above, $\underline{H}_k$ and $\underline{\ell}_k$ have the same distribution asymptotically. Hence, the same remarks apply to $\hat{x}_k^C$ and $\hat{x}_k^A$ in Eqs. (2.29) and (2.30). Therefore we may say that

$$\text{as } k \rightarrow \infty \qquad E[(\hat{x}_k^C - \hat{x}_k^A)^2] = 2\,\mathrm{Var}\,\hat{x}_k^A \tag{2.32}$$

The left hand side is the average variance of the LPT estimate. As a result the average variance of the LPT estimate asymptotically equals twice the variance of the 'learning with a teacher' estimate. This is also confirmed by the examples of Section II. 5.

II. 7. Summary

In this chapter we restrict our attention to a class of unsupervised learning problems. If the labels of the observed samples are available (or assigned) this class of problems entertains reproducing densities. We define this class of problems as Problem C.

As a solution to Problem C we suggest 'learning with a probabilistic teacher' (LPT) scheme. This scheme uses a probabilistic labelling in which the observation z_k is assigned a label generated at random with a probability density function $p_{\underline{H}_k/\underline{z}_k, \underline{\mathfrak{z}}_{k-1}, \underline{\mathcal{L}}_{k-1}}(\ell_k; z_k, \mathfrak{z}_{k-1}, \mathcal{L}_{k-1})$. The convergence properties of this scheme are established and the questions relating to the implementation are examined. Some examples are presented which show a comparison of the results of this scheme with the 'learning with a teacher' and the decision directed learning schemes. Further we find that the average asymptotic variance of the LPT estimate is twice the variance of the 'learning with a teacher' estimate.

Appendix A - The Convergence of the LPT Scheme

In this appendix we establish the convergence of the LPT scheme in the sense of Eq. (2.1). Fralick [3] has proved the convergence of the Bayesian learning scheme. Here we present a similar proof of the convergence of the LPT scheme.

First let us prove a more general theorem about the sequence of the posterior density functions $p_{\underline{x}/\underline{\mathcal{Z}}_k, \underline{\mathcal{L}}_k}(x; \mathcal{Z}_k, \mathcal{L}_k)$. This theorem has been proved in [2] for a discrete $\mathcal{H}$ space. It was proved by Daly [5] for the posterior density functions $p_{\underline{x}/\underline{\mathcal{Y}}_k}(x; \mathcal{Y}_k)$.

Theorem 1.

Any sequence $(\mathcal{G}_1, \mathcal{G}_2, \ldots)$ such that

$$\mathcal{G}_k = \int_{\mathbb{X}} f(x) p_{\underline{x}/\underline{\mathcal{Z}}_k, \underline{\mathcal{L}}_k}(x; \mathcal{Z}_k, \mathcal{L}_k) dx \tag{A-1}$$

is a bounded martingale if

(i) $f(x)$ is any non-negative Lebesgue measurable function,

(ii) $\max f(x) = M < \infty$,

(iii) $p_{\underline{x}/\underline{\mathcal{Z}}_k, \underline{\mathcal{L}}_k}(x; \mathcal{Z}_k, \mathcal{L}_k)$ is computed using the LPT scheme,

(iv) $p_{\underline{z}_{k+1}/\underline{\ell}_{k+1}, \underline{x}, \underline{\mathcal{Z}}_k, \underline{\mathcal{L}}_k}(z_{k+1}; \ell_{k+1}, x, \mathcal{Z}_k, \mathcal{L}_k) \neq 0$ for any $x \in \mathbb{X}$, $z_{k+1} \in \mathbb{Z}$, $\mathcal{L}_{k+1} \in \mathcal{H}$.

Proof.

To prove that the sequence $(\mathcal{G}_1, \mathcal{G}_2, \ldots)$ is a bounded martingale we have to show that

(a) $E\{|\mathcal{G}_k|\} < \infty$ (A-2)

(b) $E\{\mathcal{G}_{k+1}/\underline{\mathcal{Z}}_k, \underline{\mathcal{L}}_k\} = \mathcal{G}_k$ (A-3)

Since f(x) is non-negative and bounded by M on $\mathbb{X}$

$$0 \leqslant \mathcal{g}_k = \int_{\mathbb{X}} f(x) p_{\underline{x}/\underline{\mathcal{z}}_k, \underline{\mathcal{L}}_k}(x; \mathcal{z}_k, \mathcal{L}_k) dx$$

$$\leqslant M \int_{\mathbb{X}} p_{\underline{x}/\underline{\mathcal{z}}_k, \underline{\mathcal{L}}_k}(x; \mathcal{z}_k, \mathcal{L}_k) dx = M < \infty \quad . \tag{A-4}$$

Hence

$$|\mathcal{g}_k| < \infty \tag{A-5}$$

and

$$E\{|\mathcal{g}_k|\} < \infty \tag{A-6}$$

To show (b) let us evaluate $E\{\mathcal{g}_{k+1}/\underline{\mathcal{z}}_k, \underline{\mathcal{L}}_k\}$ as

$$E\{\mathcal{g}_{k+1}/\underline{\mathcal{z}}_k, \underline{\mathcal{L}}_k\} = E\{[\int_{\mathbb{X}} f(x) p_{\underline{x}/\underline{\mathcal{z}}_{k+1}, \underline{\mathcal{L}}_{k+1}}(x; \mathcal{z}_{k+1}, \mathcal{L}_{k+1}) dx]/\underline{\mathcal{z}}_k, \underline{\mathcal{L}}_k\}$$

$$= \int_{\mathbb{X}} f(x) \{E[p_{\underline{x}/\underline{\mathcal{z}}_{k+1}, \underline{\mathcal{L}}_{k+1}}(x; \mathcal{z}_{k+1}, \mathcal{L}_{k+1})/\mathcal{z}_k, \mathcal{L}_k]\} dx \tag{A-7}$$

As we can write

$$p_{\underline{x}/\underline{\mathcal{z}}_{k+1}, \underline{\mathcal{L}}_{k+1}}(x; \mathcal{z}_{k+1}, \mathcal{L}_{k+1}) =$$

$$\frac{p_{\underline{z}_{k+1}/\underline{\ell}_{k+1}, \underline{x}, \underline{\mathcal{z}}_k, \underline{\mathcal{L}}_k}(z_{k+1}; \ell_{k+1}, x, \mathcal{z}_k, \mathcal{L}_k) p_{\underline{H}_{k+1}}(\ell_{k+1})}{p_{\underline{z}_{k+1}/\underline{\ell}_{k+1}, \underline{\mathcal{z}}_k, \underline{\mathcal{L}}_k}(z_{k+1}; \ell_{k+1}, \mathcal{z}_k, \mathcal{L}_k) p_{\underline{H}_{k+1}}(\ell_{k+1})}$$

$$\cdot \; p_{\underline{x}/\underline{\mathcal{z}}_k, \underline{\mathcal{L}}_k}(x; \mathcal{z}_k, \mathcal{L}_k) \tag{A-8}$$

where

$$p_{\underline{z}_{k+1}/\underline{\ell}_{k+1},\underline{\mathcal{Z}}_k,\underline{\mathcal{L}}_k}(z_{k+1};\ell_{k+1},\mathcal{Z}_k,\mathcal{L}_k) =$$

$$\int_{\underline{X}} p_{\underline{z}_{k+1}/\underline{\ell}_{k+1},\underline{x},\underline{\mathcal{Z}}_k,\underline{\mathcal{L}}_k}(z_{k+1};\ell_{k+1},x,\mathcal{Z}_k,\mathcal{L}_k)$$

$$\cdot\ p_{\underline{x}/\underline{\mathcal{Z}}_k,\underline{\mathcal{L}}_k}(x;\mathcal{Z}_k,\mathcal{L}_k)dx \qquad \text{(A-9)}$$

if $p_{\underline{H}_{k+1}}(\ell_{k+1}) \neq 0$, we may express (A-7) as

$$E\{\mathcal{G}_{k+1}/\underline{\mathcal{Z}}_k,\underline{\mathcal{L}}_k\} = \int_{\underline{X}} f(x)p_{\underline{x}/\underline{\mathcal{Z}}_k,\underline{\mathcal{L}}_k}(x;\mathcal{Z}_k,\mathcal{L}_k)$$

$$\cdot\ E\left\{\left[\frac{p_{\underline{z}_{k+1}/\underline{\ell}_{k+1},\underline{x},\underline{\mathcal{Z}}_k,\underline{\mathcal{L}}_k}(z_{k+1};\ell_{k+1},x,\mathcal{Z}_k,\mathcal{L}_k)}{p_{\underline{z}_{k+1}/\underline{\ell}_{k+1},\underline{\mathcal{Z}}_k,\underline{\mathcal{L}}_k}(z_{k+1};\ell_{k+1},\mathcal{Z}_k,\mathcal{L}_k)}\right]\Big/\underline{\mathcal{Z}}_k,\underline{\mathcal{L}}_k\right\} dx \quad . \qquad \text{(A-10)}$$

The expectation on the right hand side may be expressed as

$$E\left\{\frac{p_{\underline{z}_{k+1}/\underline{\ell}_{k+1},\underline{x},\underline{\mathcal{Z}}_k,\underline{\mathcal{L}}_k}(z_{k+1};\ell_{k+1},x,\mathcal{Z}_k,\mathcal{L}_k)}{p_{\underline{z}_{k+1}/\underline{\ell}_{k+1},\underline{\mathcal{Z}}_k,\underline{\mathcal{L}}_k}(z_{k+1};\ell_{k+1},\mathcal{Z}_k,\mathcal{L}_k)}\Bigg/\ \underline{\mathcal{Z}}_k,\underline{\mathcal{L}}_k\right\}$$

$$= \int_{\underline{Z},\underline{H}} \frac{p_{\underline{z}_{k+1}/\underline{\ell}_{k+1},\underline{x},\underline{\mathcal{Z}}_k,\underline{\mathcal{L}}_k}(z_{k+1};\ell_{k+1},x,\mathcal{Z}_k,\mathcal{L}_k)}{p_{\underline{z}_{k+1}/\underline{\ell}_{k+1},\underline{\mathcal{Z}}_k,\underline{\mathcal{L}}_k}(z_{k+1};\ell_{k+1},\mathcal{Z}_k,\mathcal{L}_k)}$$

$$\cdot\ p_{\underline{z}_{k+1},\underline{\ell}_{k+1}/\underline{\mathcal{Z}}_k,\underline{\mathcal{L}}_k}(z_{k+1},\ell_{k+1};\mathcal{Z}_k,\mathcal{L}_k)dz_{k+1},d\ell_{k+1} \qquad \text{(A-11)}$$

$$= \int_{\mathcal{Z},\mathcal{H}} \frac{p_{\underline{z}_{k+1}/\underline{\ell}_{k+1},\underline{x},\underline{\mathfrak{z}}_k,\underline{\mathcal{L}}_k}(z_{k+1};\ell_{k+1},x,\mathfrak{z}_k,\mathcal{L}_k)}{p_{\underline{z}_{k+1}/\underline{\ell}_{k+1},\underline{\mathfrak{z}}_k,\underline{\mathcal{L}}_k}(z_{k+1};\ell_{k+1},\mathfrak{z}_k,\mathcal{L}_k)}$$

$$\cdot\ p_{\underline{\ell}_{k+1}/\underline{z}_{k+1},\underline{\mathfrak{z}}_k,\underline{\mathcal{L}}_k}(\ell_{k+1};z_{k+1},\mathfrak{z}_k,\mathcal{L}_k)$$

$$\cdot\ p_{\underline{z}_{k+1}/\underline{\mathfrak{z}}_k,\underline{\mathcal{L}}_k}(z_{k+1};\mathfrak{z}_k,\mathcal{L}_k)dz_{k+1},d\ell_{k+1}$$

$$= \int_{\mathcal{Z},\mathcal{H}} \frac{p_{\underline{z}_{k+1}/\underline{\ell}_{k+1},\underline{x},\underline{\mathfrak{z}}_k,\underline{\mathcal{L}}_k}(z_{k+1};\ell_{k+1},x,\mathfrak{z}_k,\mathcal{L}_k)}{p_{\underline{z}_{k+1}/\underline{\ell}_{k+1},\underline{\mathfrak{z}}_k,\underline{\mathcal{L}}_k}(z_{k+1};\ell_{k+1},\mathfrak{z}_k,\mathcal{L}_k)}$$

$$\cdot\ \frac{p_{\underline{z}_{k+1}/\underline{\ell}_{k+1},\underline{\mathfrak{z}}_k,\underline{\mathcal{L}}_k}(z_{k+1};\ell_{k+1},\mathfrak{z}_k,\mathcal{L}_k)p_{H_{k+1}}(\ell_{k+1})}{p_{\underline{z}_{k+1}/\underline{\mathfrak{z}}_k,\underline{\mathcal{L}}_k}(z_{k+1};\mathfrak{z}_k,\mathcal{L}_k)}$$

$$\cdot\ p_{\underline{z}_{k+1}/\underline{\mathfrak{z}}_k,\underline{\mathcal{L}}_k}(z_{k+1};\mathfrak{z}_k,\mathcal{L}_k)dz_{k+1}d\ell_{k+1} \qquad \text{(A-12)}$$

where

$$p_{\underline{z}_{k+1}/\underline{\mathfrak{z}}_k,\underline{\mathcal{L}}_k}(z_{k+1};\mathfrak{z}_k,\mathcal{L}_k) =$$

$$\int_{\mathcal{H}} p_{\underline{z}_{k+1}/\underline{\ell}_{k+1},\underline{\mathfrak{z}}_k,\underline{\mathcal{L}}_k}(z_{k+1};\ell_{k+1},\mathfrak{z}_k,\mathcal{L}_k)p_{\underline{H}_{k+1}}(\ell_{k+1})d\ell_{k+1} \qquad \text{(A-13)}$$

As, according to (iv)

$$p_{\underline{z}_{k+1}/\underline{\ell}_{k+1},\underline{x},\underline{\mathfrak{z}}_k,\underline{\mathcal{L}}_k}(z_{k+1};\ell_{k+1},x,\mathfrak{z}_k,\mathcal{L}_k) \neq 0 \quad \text{for any} \quad x \in \mathcal{X}, \quad z_{k+1} \in \mathcal{Z}, \quad \ell_{k+1} \in \mathcal{H} \qquad \text{(A-14)}$$

from (A-9)

$$p_{\underline{z}_{k+1}/\underline{\ell}_{k+1},\underline{\mathfrak{z}}_k,\underline{\mathcal{L}}_k}(z_{k+1};\ell_{k+1},\mathfrak{z}_k,\mathcal{L}_k) \neq 0 \tag{A-15}$$

and from (A-13)

$$p_{\underline{z}_{k+1}/\underline{\mathfrak{z}}_k,\underline{\mathcal{L}}_k}(z_{k+1};\mathfrak{z}_k,\mathcal{L}_k) \neq 0 \tag{A-16}$$

Under these conditions we may write the right hand side of (A-12) as

$$= \int_{\underline{Z},\underline{H}} p_{\underline{z}_{k+1}/\underline{\ell}_{k+1},\underline{x},\underline{\mathfrak{z}}_k,\underline{\mathcal{L}}_k}(z_{k+1};\ell_{k+1},x,\mathfrak{z}_k,\mathcal{L}_k)$$

$$\cdot\, p_{\underline{H}_{k+1}}(\ell_{k+1})dz_{k+1},d\ell_{k+1}$$

$$= 1 \quad . \tag{A-17}$$

Substituting this in (A-10) we get

$$E\{\mathfrak{g}_{k+1}/\underline{\mathfrak{z}}_k,\underline{\mathcal{L}}_k\} = \int_{\underline{X}} f(x)p_{\underline{x}/\underline{\mathfrak{z}}_k,\underline{\mathcal{L}}_k}(x;\mathfrak{z}_k,\mathcal{L}_k)dx$$

$$= \mathfrak{g}_k$$

This proves (b) and the theorem.

This establishes that the sequence $(\mathfrak{g}_1, \mathfrak{g}_2, \ldots)$ is a bounded martingale and as Doob [6] has shown, a bounded martingale converges with probability one. Hence the sequence $(\mathfrak{g}_1, \mathfrak{g}_2, \ldots)$ converges w. p. 1 to a value $\mathfrak{g}_\infty$, which is independent of the sequence $\{\mathfrak{z}_k, \mathcal{L}_k\}$ along which the posterior density functions are computed.

In order to show that the LPT scheme converges to the right value we use Theorem 2 which is a modified version of a theorem due to Braverman [7] and Fralick [3].

Theorem 2.

If there exists a sequence of functions $\{\phi_m(\mathfrak{z}_m, \mathcal{L}_m)\}$ defined on $\mathbb{X}$ such that $\lim_{m\to\infty} \phi_m = x_o$ with probability one, where x_o is the correct value of x, then

$$\lim_{k\to\infty} p_{\underline{x}/\underline{\mathfrak{z}}_k, \underline{\mathcal{L}}_k}(x; \mathfrak{z}_k, \mathcal{L}_k) = \delta(x - x_o) \qquad \text{w.p. } 1 \tag{A-18}$$

Proof.

Let us consider the sequence of functions

$$P_k(E_x) = \int_{E_x} p_{\underline{x}/\underline{\mathfrak{z}}_k, \underline{\mathcal{L}}_k}(x; \mathfrak{z}_k, \mathcal{L}_k)dx \tag{A-19}$$

where E_x is some set defined in $\mathbb{X}$. We can write

$$P_k(E_x) = \int_{\mathbb{X}} I_{E_x} p_{\underline{x}/\underline{\mathfrak{z}}_k, \underline{\mathcal{L}}_k}(x; \mathfrak{z}_k, \mathcal{L}_k)dx \tag{A-20}$$

where

$$I_{E_x} = \begin{cases} 1 & \text{if} \quad x \in E_x \\ 0 & \text{if} \quad x \notin E_x \end{cases}$$

is the indicator function of the set $E_x \in \mathbb{X}$. Hence from the Theorem 1 the sequence of functions $P_k(E_x)$ forms a bounded martingale and

$$\lim_{k\to\infty} P_k(E_x) = P_\infty(E_x) \qquad \text{w.p. } 1 \tag{A-21}$$

Further, if $\{\underline{u}, \underline{y}_1, \underline{y}_2, \ldots, \underline{y}_k, \ldots\}$ is a sequence of random variables such that $E[\underline{u}/\underline{y}_1, \underline{y}_2, \ldots, \underline{y}_k]$ is a bounded martingale then $E[\underline{u}/\underline{y}_1, \underline{y}_2, \ldots, \underline{y}_k]$ converges to $\underline{u}$ with probability one [8]. If we let $\underline{u} = I_{E_x}$ and $y_i = (z_i, \ell_i)$ then $E[\underline{u}/y_1, \ldots, y_k]$ becomes $P_k(E_x)$ which is a bounded martingale as shown above and hence must converge to I_{E_x} which is either 0 or 1.

Therefore $P_\infty(E_x)$ is a step function with a discontinuity at some value of x <u>which is independent of the sequence $(\mathfrak{z}_k, \underline{\mathcal{L}}_k)$.</u> But if the observed sequence was $(\mathfrak{z}_m, \mathcal{L}_m)$ on which there exists the sequence of functions $\phi_m(\mathfrak{z}_m, \mathcal{L}_m)$ converging to x_o w.p. 1, computed along this sequence the discontinuity will occur at x_o with probability one. And as $P_\infty(E_x)$ is independent of the sequence, this implies that the discontinuity of $P_\infty(E_x)$ will occur at x_o for any sequence. Hence

$$\lim_{k\to\infty} p_{\underline{x}/\mathfrak{z}_k, \underline{\mathcal{L}}_k}(x; \mathfrak{z}_k, \mathcal{L}_k) = \delta(x - x_o) \qquad \text{w.p. } 1 \tag{A-25}$$

Appendix B - An Example Violating the Assumption (A4)

Let us consider the following problem:

Let

$$p_{\underline{z}/\underline{H}}(z; H^o) = \frac{1}{2a}\,[u(z + a) - u(z - a)]$$

and

$$p_{\underline{z}/\underline{H}}(z; H^1) = \frac{1}{2b}\,[u(z + b) - u(z - b)]$$

where

$$u(\alpha) = \begin{cases} 1 & \text{if} \quad \alpha > 0 \\ 0 & \text{if} \quad \alpha \leqslant 0 \end{cases}$$

i.e. the two classes have a uniform distribution (Figure 2.7a) and

$$p_{\underline{H}}(H^o) = p$$

$$p_{\underline{H}}(H^1) = 1 - p$$

Let a be the unknown constant. We observe $\mathfrak{z}_k$ a sequence of independent, identically distributed random variables and have to 'learn' the value of the constant a.

For this problem $p_{\underline{z}_k/\underline{\ell}_k, \underline{x}, \underline{\mathfrak{z}}_{k-1}, \underline{\mathcal{L}}_{k-1}}(z_k; \ell_k, x, \mathfrak{z}_{k-1}, \mathcal{L}_{k-1})$ has the form

$$p_{\underline{z}_k/\underline{\ell}_k, \underline{x}, \underline{\mathfrak{z}}_{k-1}, \underline{\mathcal{L}}_{k-1}}(z_k; H^o, x, \mathfrak{z}_{k-1}, \mathcal{L}_{k-1}) = \frac{1}{2x}[u(z_k + x) - u(z_k - x)]$$

which is not a non-zero function. Hence the assumption (A4) is violated. It does satisfy all the other conditions required for Problem C. Spragins [10] has shown that the reproducing density function has the form

$$p_{\underline{x}/\underline{\mathfrak{z}}_k, \underline{\mathcal{L}}_k}(x; \mathfrak{z}_k, \mathcal{L}_k) = \frac{k - 1}{M_k}\left(\frac{M_k}{x}\right)^k u(x - M_k)$$

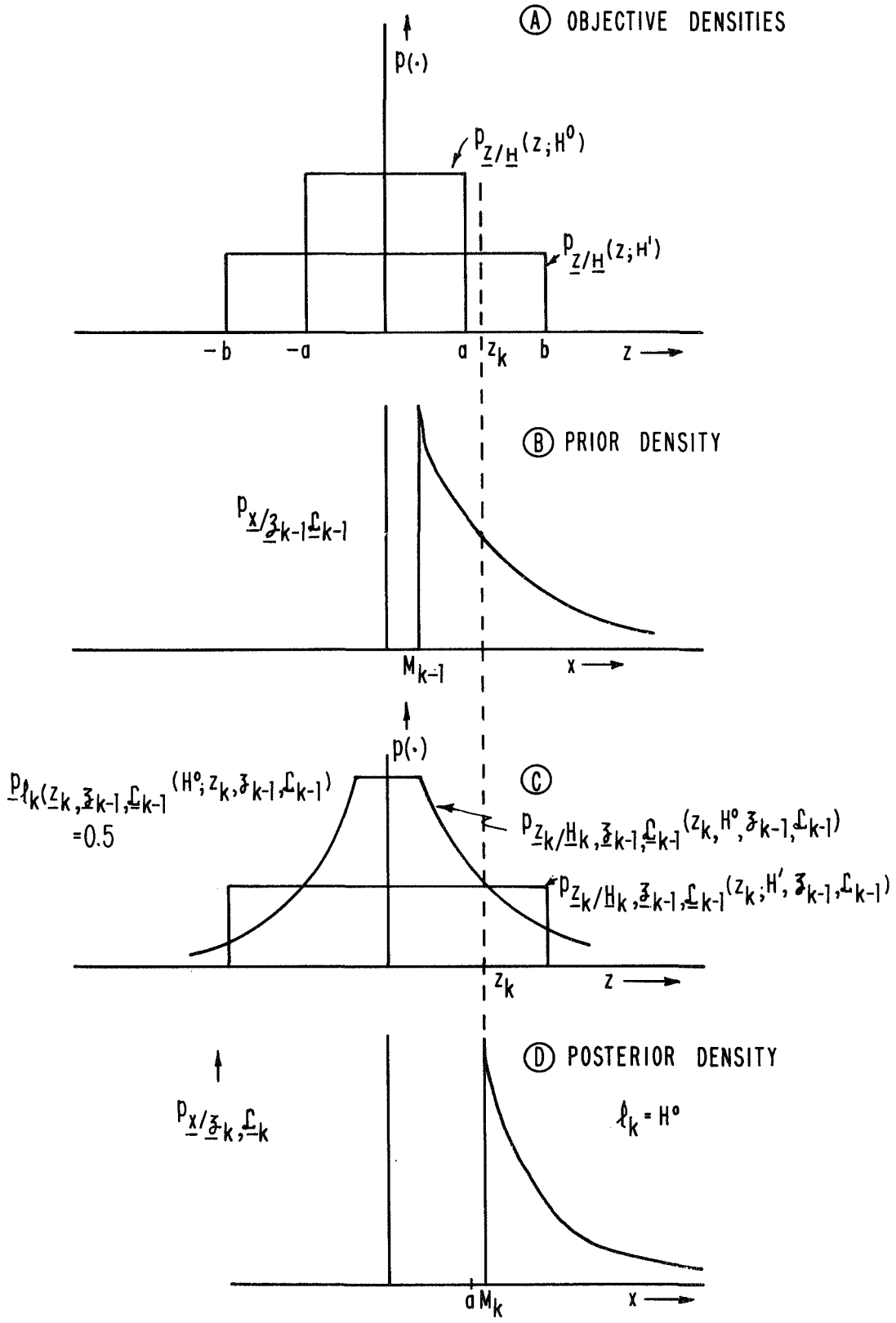

Ⓐ OBJECTIVE DENSITIES
p(·)
$p_{\underline{Z}/\underline{H}}(z;H^0)$
$p_{\underline{Z}/\underline{H}}(z;H')$
−b
−a
a
z_k
b
z
Ⓑ PRIOR DENSITY
$p_{\underline{x}/\underline{z}_{k-1}\underline{\ell}_{k-1}}$
M_{k-1}
x
p(·)
Ⓒ
$p_{\underline{\ell}_k(\underline{z}_k,\underline{z}_{k-1},\underline{\ell}_{k-1}}(H^0;z_k,\underline{z}_{k-1},\underline{\ell}_{k-1})$
=0.5
$p_{\underline{Z}_k/\underline{H}_k,\underline{z}_{k-1},\underline{\ell}_{k-1}}(z_k,H^0,\underline{z}_{k-1},\underline{\ell}_{k-1})$
$p_{\underline{Z}_k/\underline{H}_k,\underline{z}_{k-1},\underline{\ell}_{k-1}}(z_k;H',\underline{z}_{k-1},\underline{\ell}_{k-1})$
z_k
z
Ⓓ POSTERIOR DENSITY
$p_{\underline{x}/\underline{z}_k,\underline{\ell}_k}$
$\ell_k = H^0$
a M_k
x

FIGURE 2.7

where

$$m_i = \begin{cases} |z_i| & \text{if} \quad \ell_i = H^o \\ 0 & \text{if} \quad \ell_i = H^1 \end{cases}$$

and

$$M_k = \text{Max}\{m_i\}$$

Figure (2.7b) shows one such function.

If we want to use the LPT scheme for this problem, for any observed z_k we compute $p_{\underline{H}_k/\underline{z}_k, \underline{\mathcal{Z}}_{k-1}, \underline{\mathcal{L}}_{k-1}}$ using Eqs. (2.2), (2.3) and (2.4). We may select a label ℓ_k according to (2.5). Note that this assignment of the label is done at random so that if $p_{\underline{H}_k/\underline{z}_k, \underline{\mathcal{Z}}_{k-1}, \underline{\mathcal{L}}_{k-1}}(H^1; z_k, \mathcal{Z}_{k-1}, \mathcal{L}_{k-1})$ is not zero for some z_k the assigned label may be H^1. The updating may now be carried out using Eq. (2.6).

The reproducing form of the posterior density function as given above is zero till the maximum absolute value of all z_i's assigned the label H^o. If $a > b$ this will converge to the largest observed sample and in the limit converge to the correct value of a. But if $a < b$, a single assignment of a z_i, having a value greater than a (therefore it came from the class H^1) to the class H^o will rule out the correct value of the unknown constant once and for all. Fig. 2.7 shows one such sequence of computations. The observed z_k ($> a$) is from class H^1. The probability $p_{\underline{H}_k/\underline{z}_k, \underline{\mathcal{Z}}_{k-1}, \underline{\mathcal{L}}_{k-1}}(H^o; z_k, \mathcal{Z}_{k-1}, \mathcal{L}_{k-1})$ for this observation is 0.5. Therefore with a probability of 0.5, $\ell_k = H^o$. When this happens the posterior density function rules out the correct value of a at this stage. The probability of such an assignment is finite for the

LPT scheme. Therefore the LPT scheme will not give the correct result for this problem.

Appendix C - Block Diagram Structure of Solutions to Learning Problems in Pattern Classification

The basic structure of various learning schemes considered in this work can be described in terms of the following blocks

1. $\xrightarrow{H_k} \boxed{S} \xrightarrow{z_k}$. This block accepts H_k as the input and gives a z_k. This happens in the system under observation in such a way that z_k is available. H_k can be available to the "supervisor" or "teacher" only.

2. Compute Probabilities -- $\xrightarrow{z_k} \boxed{C} \xrightarrow{p_{\underline{H}_k/\underline{z}_k}}$

   This block accepts z_k as the input and computes $p_{\underline{H}_k/\underline{z}_k}$. If some x is unknown, it accepts $p_{x/\text{past info}}$ as the prior distribution of x and makes use of it in the computations.

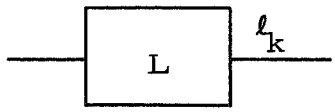
3. Bayesian Decision -- $\xrightarrow{p_{\underline{H}_k/\underline{z}_k}, \text{past info}} \boxed{D} \xrightarrow{\text{class}}$

   This block takes the bayesian decision of the class of z_k from p_{H_k/z_k}, past info.

4. Teacher -- $\xrightarrow{H_k} \boxed{T} \xrightarrow{H_k}$

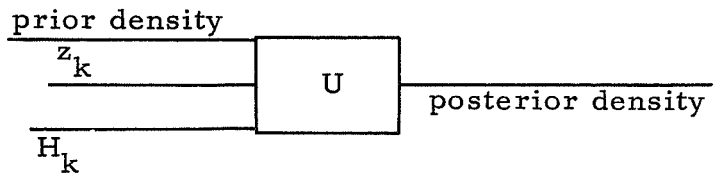
   This block has access to H_k, the correct class of H_k. It makes H_k available at its output.

5. Labelling --



This block performs labelling.

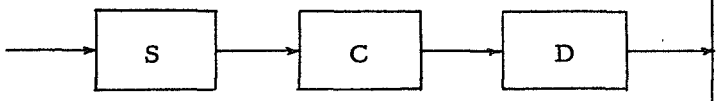
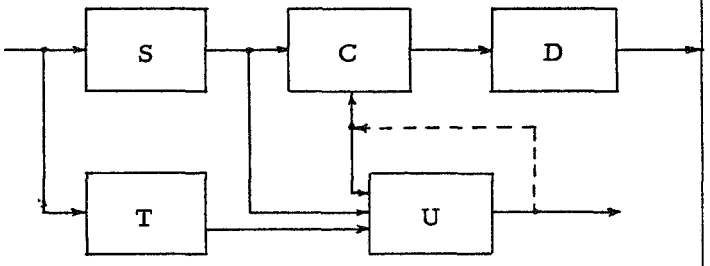
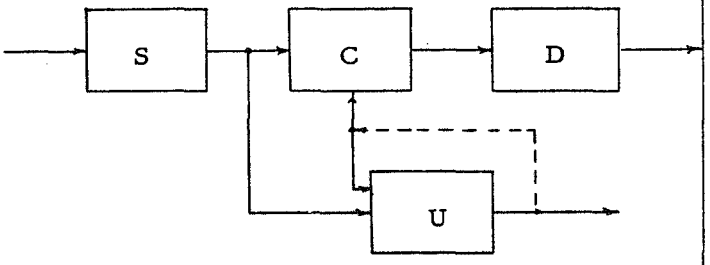
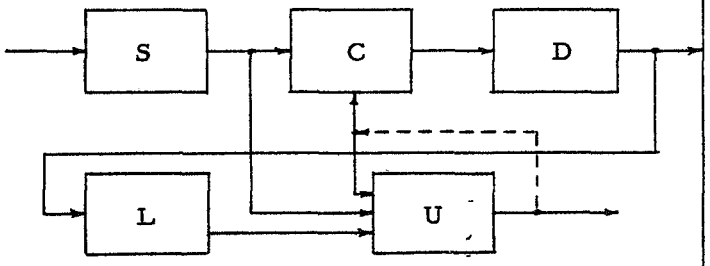
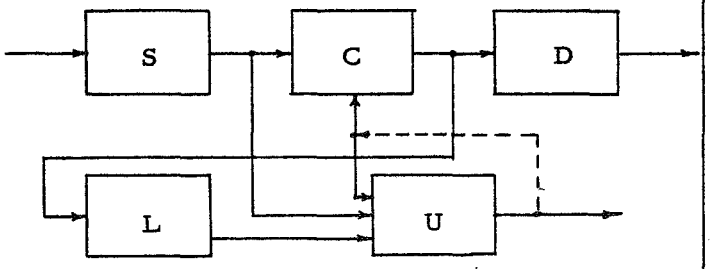
6. Updating --



This block accepts the prior density, z_k (and H_k or ℓ_k if available) and computes the posterior density.

Based on these blocks, the structure of various schemes is presented in Table 1.

TABLE 1

	SCHEME	BLOCK DIAGRAM
1	CLASSIFICATION	
2	LEARNING WITH A TEACHER	
3	LEARNING WITHOUT A TEACHER	
4	DECISION DIRECTED LEARNING	
5	LEARNING WITH A PROBABILISTIC TEACHER	

References for Chapter II

1. Agrawala, A. K., Ho, Y. C., Wackerbarth, R. K., "Learning Without a Teacher with Reproducing Densities," Nerem Record, Vol. 10, pp. 102-103, 1968.

2. Agrawala, A. K., "Learning With A Probabilistic Teacher," IEEE Trans. On Info. Th., Vol. IT-16, No. 3, July 1970.

3. Fralick, S. C., "Learning to Recognize Patterns Without A Teacher," Stanford University, Stanford, Calif., Tech. Rept. 6103-10, March 1965.

4. Hull, T. E. and Dobell, A. R., "Random Number Generators," SIAM Rev., Vol. 4, pp. 230-254, July 1962.

5. Daly, R. F., "The Adaptive Binary Detection Problem on the Real Line," Stanford University, Stanford, Calif., Tech. Rept. 2003-3, Feb. 1962.

6. Doob, J. L., 'Stochastic Processes,' Wiley and Sons, New York, 1953.

7. Braverman, D., "Machine Learning and Automatic Pattern Recognition," Stanford University, Stanford, Calif., Tech. Rept. 2003-1, Feb. 1961.

8. Breiman, L., 'Probability,' Addison Wesley, Reading, Mass., 1968. (Theorem 5.21, Corr. 5.22, Pages 92-93.)

9. Ralston, A., 'A First Course in Numerical Analysis,' McGraw Hill, New York, 1965.

10. Spragins, J. D., "Reproducing Distributions for Machine Learning," Stanford University, Stanford, Calif., Tech. Rept. 6103-7, Nov. 1963.

11. Teicher, H., "On the Mixtures of Distributions," Ann. of Math. Stat., Vol. 31, pp. 55-73, 1960.

12. Raiffa, H., and Schlaifer, R., 'Applied Statistical Decision Theory,' Harvard Business School, Boston, 1961.

CHAPTER III

APPLICATION OF THE LPT SCHEME TO GAUSS MARKOV SEQUENCE

In all the estimation problems we have considered so far in this work we wanted to estimate the unknown value of a parameter. We were allowed to observe a sequence of "samples" which may or may not contain information about the parameter. The unknown value of the parameter remained fixed as we observed the sequence of samples. Therefore we observed a sequence of independent identically distributed samples.

In various practical problems the parameter of interest is the state of some dynamic system. For example we may be interested in tracking the position of a satellite. The state of such a dynamic system and hence the "parameter" we want to estimate, keeps on changing.

Let the dynamics of the system be defined by

$$\underline{x}_{k+1} = \phi_k \underline{x}_k + \Gamma_k \underline{w}_k$$

where $\underline{w}_k$ is a gaussian purely random sequence of known mean and variance. ϕ_k and Γ_k are known constants. $\underline{x}_k$ is the state of the system at the kth stage. We note that $\underline{x}_k$, $k = 1, 2, 3, \ldots$ is a Gauss Markov sequence. We are interested in estimating x_k the state of this Gauss Markov sequence at the kth stage.

We may adopt the Bayesian estimation philosophy for this estimation. In that case some a priori knowledge of the state is required. This may be in terms of the prior distribution for $\underline{x}_1$. From this prior

knowledge we can make the estimates of the states for all k. But if we are allowed to observe some samples we may be able to improve these estimates. If the state could be observed directly we can make the perfect estimate. In practice the state of the system can be observed but not directly or perfectly. At the kth stage we can observe a z_k which is some function of the state x_k. In addition z_k may contain some observation noise also. This observation can be used to improve the estimate made from the prior information.

Let us consider the observation z_k as an outcome of $\underline{z}_k$* where

$$\underline{z}_k = H_k \underline{x}_k + \underline{v}_k$$

and $\underline{v}_k$ is a gaussian zero mean independent white sequence. When H_k is a known constant, the Bayesian estimation of $\underline{x}_k$ making use of z_k is like a supervised learning problem (Problem A) of Chapter I and can be solved rather easily in terms of the well known Kalman-Bucy filter [4].

In some practical problems H_k cannot be considered as a constant. For example in the tracking of a satellite some observations do not contain the signal which results in $\underline{H}_k$, having a value 0 or H_k, at random. When we consider $\underline{H}_k$ as a random variable the estimation problem becomes an unsupervised estimation problem. As we have seen in Chapter I the Bayesian estimation leads to an infeasible solution to this problem.

In Chapter II we noted that the LPT scheme leads to a feasible solution of the unsupervised learning problem. Here we show how the

---

* We are treating the variables as scalars in this chapter. The discussion is valid if x_k and z_k are vectors and ϕ_k, Γ_k and H_k appropriate matrices.

LPT scheme can be used to estimate the state of a Gauss Markov sequence.

The only work reported in literature on similar problems is by Nahi [1]. He has constructed the best linear estimate for the problem with $\underline{H}_k$ having a binary distribution. The inherent nonlinearity of the problem suggests that some nonlinear estimate may be better than the best linear estimate.

In this chapter we define a Problem D in which the unknown parameter x_k forms a Gauss Markov sequence. The observation process is defined so that the estimation problem is an unsupervised learning problem. We formulate a solution to Problem D using the LPT scheme. An example is presented in which we compare the performance of the LPT scheme solution to the best linear solution of Nahi.

III.1. Problem Formulation - Problem D

Let us define Problem D as follows:

Consider a discrete Gauss Markov Process $\underline{x}_k$, $k = 1, 2, 3, \ldots$ defined by

$$\underline{x}_{k+1} = \phi_k \underline{x}_k + \Gamma_k \underline{w}_k \tag{3.1}$$

ϕ_k and Γ_k are known. $\underline{w}_k$ is a Gaussian white noise sequence such that

$$E\{\underline{w}_k\} = \overline{w}_k \tag{3.2}$$

$$E\{(\underline{w}_{k_1} - \overline{w}_{k_1})(\underline{w}_{k_2} - \overline{w}_{k_2})\} = Q_{k_1} \cdot \delta_{k_1 k_2}{}^{*} \tag{3.3}$$

---

* $\delta_{k_1 k_2}$ is the kroneker delta function, i.e.

$$\delta_{k_1 k_2} = \begin{cases} 1 & \text{if} \quad k_1 = k_2 \\ 0 & \text{if} \quad k_1 \neq k_2 \end{cases}$$

We observe $\underline{z}_k$ where

$$\underline{z}_k = \underline{H}_k \underline{x}_k + \underline{v}_k \tag{3.4}$$

The observation noise $\underline{v}_k$ is an independent gaussian white noise sequence and

$$E\{\underline{v}_k\} = 0 \tag{3.5}$$

$$E\{\underline{v}_{k_1} v_{k_2}\} = R_{k_1} \cdot \delta_{k_1 k_2} \tag{3.6}$$

In addition to the additive noise $\underline{v}_k$ the observation $\underline{z}_k$ has a multiplicative noise $\underline{H}_k$. $\underline{H}_k$ is a random variable independent of all other random variables and has the probability density function $p_{\underline{H}_k}(H_k)$. This density function is known for all k.

We are given $p_{\underline{x}_1}(x_1) \sim N(\bar{x}_1, M_1)$ as the prior density function for $\underline{x}_1$. At the kth stage after observing a sequence $\mathfrak{z}_k$ we have to make an estimate of $\underline{x}_k$ as $\hat{x}_k$. In Section I. 3 we have seen that it is sufficient to compute the posterior density functions to make such optimal estimates. Here we accept the posterior density functions as a solution.*

We note that Problem D is an extension of Problem B-1 (of Chapter I) and Problem C-1 (of Chapter II). In all these problems the observation process is the same. In B-1 and C-1 x was an unknown constant. Its value was unknown but fixed for the observation sequence. For Problem D we allow the value of x to change as a Gauss Markov sequence.

---

*Deutsch [2] has shown that the mean of the posterior density function is optimal for a general class of cost functions. Therefore when we have to make an estimate we shall use the mean of the posterior density function as the estimate.

III. 2. General Solution

To solve Problem D we have to compute the posterior density function $p_{\underline{x}_k/\mathcal{Z}_k}(x_k;\mathcal{Z}_k)$. If we arrange the computations sequentially, at the beginning of the kth stage (i.e. at the end of the $k-1^{st}$ stage) we have the density function $p_{\underline{x}_{k-1}/\mathcal{Z}_{k-1}}(x_{k-1};\mathcal{Z}_{k-1})$. We observe z_k and have to compute the density function $p_{\underline{x}_k/\mathcal{Z}_k}(x_k;\mathcal{Z}_k)$. This can be carried out in two steps.

(a) Dynamic Propagation - Using the system equation (3.1) and the density functions $p_{\underline{x}_{k-1}/\mathcal{Z}_{k-1}}(x_{k-1};\mathcal{Z}_{k-1})$ and $p_{\underline{w}_{k-1}}(w_{k-1})$ we compute $p_{\underline{x}_k/\mathcal{Z}_{k-1}}(x_k;\mathcal{Z}_{k-1})$.

This is a straightforward computation. As Eq. (3.1) shows, $\underline{x}_k$ is a weighted sum of two independent random variables $\underline{x}_{k-1}$ and $\underline{w}_{k-1}$. When x_k and z_k are scalars* the probability density function $p_{\underline{x}_k/\mathcal{Z}_{k-1}}$ can be computed as follows [3].

$$p_{\underline{x}_k/\mathcal{Z}_{k-1}}(x_k;\mathcal{Z}_{k-1}) = \int_{-\infty}^{\infty} \frac{1}{|\phi_{k-1}|} p_{\underline{x}_{k-1}/\mathcal{Z}_{k-1}}\left(\frac{x_k - \xi}{\phi_{k-1}};\mathcal{Z}_{k-1}\right) \cdot \frac{1}{|\Gamma_{k-1}|} p_{\underline{w}_k}\left(\frac{\xi}{\Gamma_{k-1}}\right) d\xi \qquad (3.7)$$

The right hand side of equation (3.7) is a convolution integral and can be computed for any density function in principle, and easily for gaussian distributions.

In Problem D the density function $p_{\underline{w}_k}(w_k)$ is gaussian. If $p_{\underline{x}_{k-1}/\mathcal{Z}_{k-1}}$ is also gaussian the computations of this step become

---

*If x_k and z_k are vectors this computation can still be carried out. It is rather simple for multinomial distributions [4].

very simple. The resulting $p_{\underline{x}_k/\mathcal{Z}_{k-1}}$ is also gaussian and

$$p_{\underline{x}_{k-1}/\mathcal{Z}_{k-1}} \sim N(\hat{x}_{k-1}, P_{k-1}) \tag{3.9}$$

and

$$p_{\underline{x}_k/\mathcal{Z}_{k-1}} \sim N(\bar{x}_k, M_k) \tag{3.10}$$

where

$$\bar{x}_k = \phi_{k-1}\hat{x}_{k-1} + \Gamma_{k-1}\bar{w}_{k-1} \tag{3.11}$$

and

$$M_k = \phi_{k-1}P_{k-1}\phi_{k-1}^T + \Gamma_{k-1}Q_{k-1}\Gamma_{k-1}^T \tag{3.12}$$

(b) Static Updating - Given the probability density function

$p_{\underline{x}_k/\mathcal{Z}_{k-1}}(x_k; \mathcal{Z}_{k-1})$ and the observation z_k
we compute the posterior density function

$p_{\underline{x}_k/\mathcal{Z}_k}(x_k; \mathcal{Z}_k)$.

We note that as $\underline{H}_k$ is a random variable this updating is the same as a single step of the unsupervised learning problem, Problem B, of Chapter I. We may use Eqs. (1.16), (1.17) and (1.18) to carry out this updating. But as we have seen in Section I.4 no reproducing densities are known to exist for this problem and the updating, therefore, requires that a complete nonparametric density function be stored and manipulated. As a result the Bayesian estimation procedure is infeasible for this step.

We note however, that if H_k is known the reproducing density functions do exist. A gaussian prior density function results in a gaussian posterior density. If

$$p_{\underline{x}_k/\mathcal{Z}_{k-1}}(x_k; \mathcal{Z}_{k-1}) \sim N(\bar{x}_k, M_k) \tag{3.10}$$

then

$$p_{\underline{x}_k/\underline{z}_k}(x_k; z_k) \sim N(\hat{x}_k, P_k) \tag{3.13}$$

where

$$\hat{x}_k = \bar{x}_k + P_k H_k^T R_k^{-1}(z_k - H_k \bar{x}_k) \tag{3.14}$$

and

$$P_k^{-1} = M_k^{-1} + H_k^T R_k^{-1} H_k \tag{3.15}$$

Eqs. (3.11), (3.12), (3.14) and (3.15) define the well known Kalman filter [4].

In Chapter II we have seen that the LPT scheme offers a solution to the unsupervised learning problem of the type considered in step (b) of Problem D. All the conditions required for the LPT scheme in Chapter II are satisfied by the problem of this step. Let us see how the LPT scheme can be used for this problem.

III. 3. The LPT Solution

We have seen above in Section III. 2 that if H_k is known, the step (b) computations become very simple. If the LPT scheme is used for step (b) we generate a label ℓ_k and treat it as the correct value of the random variable $\underline{H}_k$ which was active for the kth observation. And as we are treating H_k as known (ℓ_k), the posterior density functions at any stage remain gaussian. However, a label ℓ_k has to be generated before we can carry out the static updating of step (b). Therefore at the kth stage we may proceed as follows:

Step (a) - We are given the probability densities $p_{\underline{x}_{k-1}/\underline{z}_{k-1},\underline{\ell}_{k-1}}(x_{k-1}; z_{k-1}, \ell_{k-1})$ and $p_{\underline{w}_{k-1}}(w_{k-1})$.

Using Eq. (3.1) we compute $p_{\underline{x}_k/\underline{\mathcal{Z}}_{k-1},\underline{\mathcal{L}}_{k-1}}(x_k;\mathcal{Z}_{k-1},\mathcal{L}_{k-1})$ as

$$p_{\underline{x}_{k-1}/\underline{\mathcal{Z}}_{k-1},\underline{\mathcal{L}}_{k-1}}(x_{k-1};\mathcal{Z}_{k-1},\mathcal{L}_{k-1}) \sim N(\hat{x}_{k-1}, P_{k-1}) \tag{3.9a}$$

$$p_{\underline{x}_k/\underline{\mathcal{Z}}_{k-1},\underline{\mathcal{L}}_{k-1}}(x_k;\mathcal{Z}_{k-1},\mathcal{L}_{k-1}) \sim N(\bar{x}_k, M_k) \tag{3.10a}$$

where

$$\bar{x}_k = \phi_{k-1}\hat{x}_{k-1} + \Gamma_{k-1}\bar{w}_{k-1} \tag{3.11a}$$

and

$$M_k = \phi_{k-1}P_{k-1}\phi_{k-1}^T + \Gamma_{k-1}Q_{k-1}\Gamma_{k-1}^T \tag{3.12a}$$

Note that Eqs. (3.11a) and (3.12a) are exactly the same as Eqs. (3.11) and (3.12) respectively.

Step (b_1) - Probabilistic Labelling - Having computed $p_{\underline{x}_k/\underline{\mathcal{Z}}_{k-1},\underline{\mathcal{L}}_{k-1}}(x_k;\mathcal{Z}_{k-1},\mathcal{L}_{k-1})$ we observe z_k and proceed to generate a label ℓ_k for it.

For this $p_{\underline{H}_k/\underline{z}_k,\underline{\mathcal{Z}}_{k-1},\underline{\mathcal{L}}_{k-1}}(H_k;z_k,\mathcal{Z}_{k-1},\mathcal{L}_{k-1})$ is computed as follows. We write

$$p_{\underline{H}_k/\underline{z}_k,\underline{\mathcal{Z}}_{k-1},\underline{\mathcal{L}}_{k-1}}(H_k;z_k,\mathcal{Z}_{k-1},\mathcal{L}_{k-1}) = \frac{p_{\underline{z}_k,\underline{H}_k/\underline{\mathcal{Z}}_{k-1},\underline{\mathcal{L}}_{k-1}}(z_k,H_k;\mathcal{Z}_{k-1},\mathcal{L}_{k-1})}{p_{\underline{z}_k/\underline{\mathcal{Z}}_{k-1},\underline{\mathcal{L}}_{k-1}}(z_k;\mathcal{Z}_{k-1},\mathcal{L}_{k-1})} \tag{3.16}$$

As H_k is assumed independent random variable, we can write

$$p_{\underline{H}_k/\underline{z}_k,\underline{\mathcal{Z}}_{k-1},\underline{\mathcal{L}}_{k-1}}(H_k;z_k,\mathcal{Z}_{k-1},\mathcal{L}_{k-1}) = \frac{p_{\underline{z}_k/\underline{H}_k,\underline{\mathcal{Z}}_{k-1},\underline{\mathcal{L}}_{k-1}}(z_k;H_k,\mathcal{Z}_{k-1},\mathcal{L}_{k-1})p_{\underline{H}_k}(H_k)}{p_{\underline{z}_k/\underline{\mathcal{Z}}_{k-1},\underline{\mathcal{L}}_{k-1}}(z_k;\mathcal{Z}_{k-1},\mathcal{L}_{k-1})} \tag{3.16a}$$

Here

$$p_{\underline{z}_k/\underline{\mathfrak{z}}_{k-1},\underline{\mathcal{L}}_{k-1}}(z_k;\mathfrak{z}_{k-1},\mathcal{L}_{k-1}) =$$

$$\int_{\mathcal{H}_k} p_{\underline{z}_k/\underline{H}_k,\underline{\mathfrak{z}}_{k-1},\underline{\mathcal{L}}_{k-1}}(z_k;H_k,\mathfrak{z}_{k-1},\mathcal{L}_{k-1})p_{\underline{H}_k}(H_k)dH_k \tag{3.17}$$

and

$$p_{\underline{z}_k/\underline{H}_k,\underline{\mathfrak{z}}_{k-1},\underline{\mathcal{L}}_{k-1}}(z_k;H_k,\mathfrak{z}_{k-1},\mathcal{L}_{k-1}) =$$

$$\int_{\underline{X}_k} p_{\underline{z}_k/\underline{H}_k,\underline{x}_k,\underline{\mathfrak{z}}_{k-1},\underline{\mathcal{L}}_{k-1}}(z_k;H_k,x,\mathfrak{z}_{k-1},\mathcal{L}_{k-1})$$

$$\cdot\ p_{\underline{x}_k/\underline{H}_k,\underline{\mathfrak{z}}_{k-1},\underline{\mathcal{L}}_{k-1}}(x_k;H_k,\mathfrak{z}_{k-1},\mathcal{L}_{k-1})dx_k \quad . \tag{3.18}$$

In the way the problem is formulated, we have

$$p_{\underline{z}_k/\underline{H}_k,\underline{x}_k,\underline{\mathfrak{z}}_{k-1},\underline{\mathcal{L}}_{k-1}}(z_k;H_k,x_k,\mathfrak{z}_{k-1},\mathcal{L}_{k-1}) =$$

$$p_{\underline{z}_k/\underline{H}_k,\underline{x}_k}(z_k;H_k,x_k) \quad .$$

i.e. $\underline{z}_k$ is independent of $\mathfrak{z}_{k-1}$ and $\mathcal{L}_{k-1}$, given x_k and H_k. Also $\underline{x}_k$ is independent of $\underline{H}_k$. Hence we may write

$$p_{\underline{z}_k/\underline{H}_k,\underline{\mathfrak{z}}_{k-1},\underline{\mathcal{L}}_{k-1}}(z_k;H_k,\mathfrak{z}_{k-1},\mathcal{L}_{k-1}) =$$

$$\int_{\underline{X}_k} p_{\underline{z}_k/\underline{H}_k,\underline{x}_k}(z_k;H_k,x_k)p_{\underline{x}_k/\underline{\mathfrak{z}}_{k-1},\underline{\mathcal{L}}_{k-1}}(x_k;\mathfrak{z}_{k-1},\mathcal{L}_{k-1})dx_k \quad . \tag{3.19}$$

In generating the label ℓ_k it is treated as a random variable $\underline{\ell}_k$ which has a probability density function

$$p_{\underline{\ell}_k/\underline{z}_k,\underline{\mathfrak{z}}_{k-1},\underline{\mathcal{L}}_{k-1}}(\ell_k; z_k, \mathfrak{z}_{k-1}, \mathcal{L}_{k-1}) =$$

$$p_{\underline{H}_k/\underline{z}_k,\underline{\mathfrak{z}}_{k-1},\underline{\mathcal{L}}_{k-1}}(\ell_k; z_k, \mathfrak{z}_{k-1}, \mathcal{L}_{k-1}) \tag{3.20}$$

To generate the label we draw a random number from the probability density function $p_{\underline{\ell}_k/\underline{z}_k,\underline{\mathfrak{z}}_{k-1},\underline{\mathcal{L}}_{k-1}}(\ell_k; z_k, \mathfrak{z}_{k-1}, \mathcal{L}_{k-1})$.

Step (b_2) - Using the prior density function $p_{\underline{x}_k/\underline{\mathfrak{z}}_{k-1},\underline{\mathcal{L}}_{k-1}}(x_k; \mathfrak{z}_{k-1}, \mathcal{L}_{k-1})$, the observation z_k and the label ℓ_k, we compute the posterior density function $p_{\underline{x}_k/\underline{\mathfrak{z}}_k,\underline{\mathcal{L}}_k}(x_k; \mathfrak{z}_k, \mathcal{L}_k)$ as

$$p_{\underline{x}_k/\underline{\mathfrak{z}}_k,\underline{\mathcal{L}}_k}(x_k; \mathfrak{z}_k, \mathcal{L}_k) \sim N(\hat{x}_k, P_k) \tag{3.13a}$$

where

$$\hat{x}_k = \bar{x}_k + P_k \ell_k^T R_k^{-1}(z_k - \ell_k x_k^T) \tag{3.14a}$$

and

$$P_k^{-1} = M_k^{-1} + \ell_k^T R_k^{-1} \ell_k \quad . \tag{3.15a}$$

Figure 3.1 shows a block diagram for these computations. A knowledge of $p_{\underline{x}_1}(x_1)$ is required to start these computations.

In Problem D we are interested in estimating $\underline{x}_k$ which is a random variable and changes as a Gauss Markov sequence from stage to stage. Therefore we cannot really talk about the convergence of any scheme in terms of Eq. (2.1) (i.e. the posterior density function converging to a delta function). In this case we say that <u>a scheme leads to a converging solution if the variance of the estimate remains finite</u>.

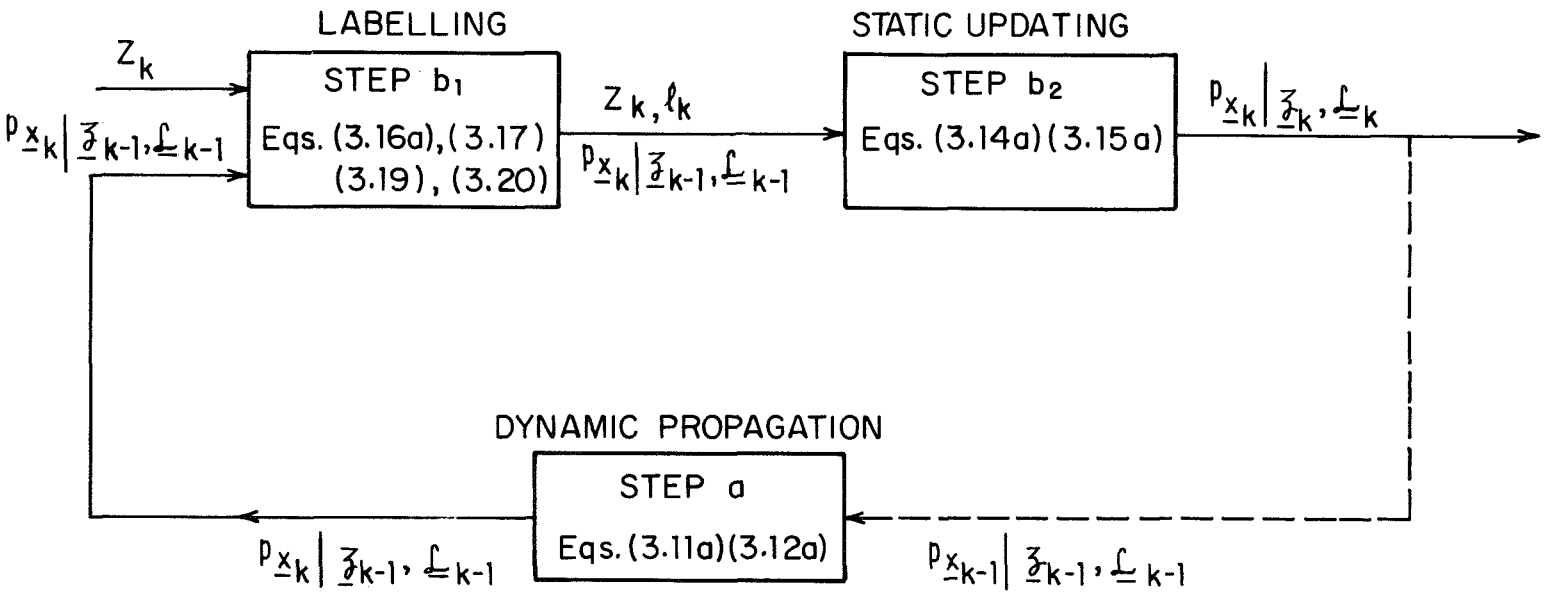


FIG. 3-1 THE LPT SOLUTION OF PROBLEM D.

If Eq. (3.1) represents a stable dynamic system, the variance of $\underline{x}_k$ attains a finite constant value for large k. From Eq. (3.15a) we note that P_k is either less than or equal to M_k. Therefore when the randomness of $\underline{\ell}_k$ is taken into account, the variance of $\hat{x}_k$ will be less than (or equal to) M_k, the variance of $\bar{x}_k$. As M_k remains finite if we make no observations, we conclude that the variance of the estimate $\hat{x}_k$ computed using the LPT scheme remains finite. Hence the LPT scheme leads to a converging solution of Problem D.

In the LPT scheme solution of Problem D as discussed above, we generate a label ℓ_k for the observation z_k and treat it as the correct value for the unknown value of the random variable $\underline{H}_k$. Hence if an estimate can be made for some system with H_k known (i.e. the system is observable [4]) the estimation can be carried out using the LPT scheme.

III. 4. The Implementation of the LPT Solution

We have to carry out the steps (a), (b_1) and (b_2) at any stage to implement the LPT solution of Section III. 3 for Problem D

The computations of step (a) and (b_2) are exactly the same as those of a Kalman filter [4]. The only difference is that the sequence ℓ_k is not known in advance and has to be generated at every stage after the sample value is observed. If H_k is known, in a practical implementation the ℓ_k sequence will have to be stored somewhere and the value H_k read in at the kth stage. In the present implementation the corresponding value, ℓ_k, is provided by the computations of step (b_1). Therefore the implementation of step (a) and (b_2) is no different from the implementation of a Kalman filter. This implementation is simple for a vector x_k and z_k case also [4].

In step (b_1) we have to generate a label ℓ_k which is used in the step (b_2) as the correct value of the random variable $\underline{H}_k$. The label ℓ_k is generated after observing z_k and as an outcome of a random variable $\underline{\ell}_k$. The random variable $\underline{\ell}_k$ has a probability density which depends on z_k. Therefore this probability density function has to be computed first. This can be done easily using Eqs. (3.16), (3.17), (3.19) and (3.20). As a result of this computation we get the probability density function $p_{\underline{\ell}_k/\underline{z}_k, \underline{\mathfrak{z}}_{k-1}, \underline{\mathcal{L}}_{k-1}}(\ell_k; z_k, \mathfrak{z}_{k-1}, \mathcal{L}_{k-1})$. We still have to generate a random outcome from $\mathcal{H}_k$ space with this distribution.

In generating such random outcomes on a general purpose digital computer we make use of a pseudo random number generator which gives a random number $\underline{\omega}$ having a uniform probability density on Ω. For a scalar ℓ_k we may take $\Omega = [0, 1]$ and use the methods discussed in Section II. 4 to generate ℓ_k.

When x_k and z_k are multidimensional vectors, $\mathcal{H}_k$ becomes a multidimensional space. If this space is discrete, i. e. if $\underline{H}$ has a finite number of allowed values only, the generation of ℓ_k from the pseudo random number is simple and it has been discussed in Section II. 4.

To consider the generation of $\underline{\ell}$ when $\mathcal{H}$ is a multidimensional continuous space, let ℓ^i be the ith component of ℓ which is considered as an n dimensional vector. Let $\underline{\ell}^i$ be defined on the real line. We know the density $p_{\underline{\ell}}(\ell) = p_{\underline{\ell}^1, \underline{\ell}^2, \underline{\ell}^3, \ldots, \underline{\ell}^n}(\ell^1, \ell^2, \ldots, \ell^n)$. The marginal density function for $\underline{\ell}^1$ is computed as

$$p_{\underline{\ell}^1}(\ell^1) = \int\int\cdots\int p_{\underline{\ell}^1, \underline{\ell}^2, \underline{\ell}^3, \ldots, \underline{\ell}^n}(\ell^1, \ell^2, \ell^3, \ldots, \ell^n) d\ell^2, d\ell^3, \ldots, d\ell^n. \tag{3.21}$$

Now $\underline{\ell}^1$ may be treated as a scalar random variable with known density function $p_{\underline{\ell}^1}(\ell^1)$ and its value may be generated using the methods of Section II. 4. Having generated a value for $\underline{\ell}^1$ as equal to ℓ^1 we compute $p_{\underline{\ell}^2/\underline{\ell}^1}(\ell^2; \ell^1)$ as

$$p_{\underline{\ell}^2/\underline{\ell}^1}(\ell^2; \ell^1) = \frac{p_{\underline{\ell}^1, \underline{\ell}^2}(\ell^1, \ell^2)}{p_{\underline{\ell}^1}(\ell^1)}$$

$$= \frac{\int\int\cdots\int p_{\underline{\ell}^1, \underline{\ell}^2, \underline{\ell}^3, \ldots, \underline{\ell}^n}(\ell^1, \ell^2, \ell^3, \ldots, \ell^n) d\ell^3, d\ell^4, \ldots, d\ell^n}{p_{\underline{\ell}^1}(\ell^1)} \tag{3.22}$$

A value for $\underline{\ell}^2$ may now be drawn from $p_{\underline{\ell}^2/\underline{\ell}^1}(\ell^2; \ell^1)$ and we may proceed to compute $p_{\underline{\ell}^3/\underline{\ell}^1, \underline{\ell}^2}$, and this may go on till we have generated the complete vector $\underline{\ell}$.

In principle the above computation can always be carried out. It is very difficult however. But we note that it arises only when $\mathcal{H}$ is a multidimensional continuous space. In this case Problem D in itself is so difficult that above may still be an attractive solution.

Let us consider an example next.

III. 5. Example

In defining Problem D in Section III. 1 we have specified the form of all the density functions except $p_{\underline{H}_k}(H_k)$. In the subsequent discussion we assumed that the form of this density function is known, though we

did not require the exact form. As an example let us consider a specific form of $p_{\underline{H}_k}$ and define Problem (D-1) as:

Problem (D-1) - Let a scalar $\underline{H}_k$ have a binary distribution, i.e.

$$\underline{H}_k = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1 - p \end{cases} \tag{3.23}$$

We consider a Problem D with a scalar $\underline{x}_k$ and $\underline{z}_k$ and $\underline{H}_k$ with above binary distribution, as Problem (D-1).

The LPT solution to this problem uses Eqs. (3.11a) and (3.12a) for step (a) and Eqs. (3.14a) and (3.15a) for step (b_2). For the labelling in step (b_1) we use Eqs. (3.16a), (3.17), (3.19) and (3.20). When H_k has a binary distribution the labelling step is exactly the same as for Problem (C-1) of Chapter II. We may then write $p_{\underline{\ell}_k/\underline{z}_k,\underline{\mathfrak{z}}_{k-1},\underline{\mathcal{L}}_{k-1}}(\ell_k; z_k,\mathfrak{z}_{k-1},\mathcal{L}_{k-1})$ as

$$p_{\underline{\ell}_k/\underline{z}_k,\underline{\mathfrak{z}}_{k-1},\underline{\mathcal{L}}_{k-1}}(1; z_k,\mathfrak{z}_{k-1},\mathcal{L}_{k-1}) =$$

$$p_{\underline{H}_k/\underline{z}_k,\underline{\mathfrak{z}}_{k-1},\underline{\mathcal{L}}_{k-1}}(1; z_k,\mathfrak{z}_{k-1},\mathcal{L}_{k-1})$$

$$= \frac{\dfrac{p}{\sqrt{2\pi(R_k + M_k)}}\, e^{-\frac{(z_k-\bar{x}_k)^2}{2(R_k+M_k)}}}{\dfrac{p}{\sqrt{2\pi(R_k+M_k)}}\, e^{-\frac{(z_k-\bar{x}_k)^2}{2(R_k+M_k)}} + \dfrac{(1-p)}{\sqrt{2\pi R_k}}\, e^{-\frac{z_k^2}{2R_k}}}$$

$$= \alpha_k \quad \text{say} \tag{3.24}$$

and

$$p_{\underline{\ell}_k/\underline{z}_k, \underline{\mathfrak{z}}_{k-1}, \underline{\mathcal{L}}_{k-1}}(0; z_k, \mathfrak{z}_{k-1}, \mathcal{L}_{k-1}) = 1 - \alpha_k \quad .$$

To generate the label we make use of a pseudo random number ω having a uniform density on $\Omega = [0, 1]$ and assign the label ℓ_k as

$$\begin{aligned} \omega &\leq \alpha_k \qquad & \ell_k &= 1 \\ \omega &> \alpha_k \qquad & \ell_k &= 0 \end{aligned} \tag{3.25}$$

This label is used in Eqs. (3.14a) and (3.15a) and we get $\hat{x}_k$ as the estimate at the kth stage.

III.5.1. The Best Linear Filter for Problem (D-1) (Nahi's Solution)

Nahi [1] has suggested a method of constructing the best linear estimate.* If $Q_k = Q$, $R_k = R$, $\overline{w}_k = 0$, $\phi_k = \phi$ and $\Gamma_k = \Gamma$ in Problem (D-1), the estimate $\hat{x}^N_{k+1}$ can be computed as follows:

$$\hat{x}^N_{k+1} = F_{1k}\hat{x}^N_k + F_{2k}z_k \tag{3.26}$$

where

$$F_{1k} = \phi - pF_{2k} \tag{3.27}$$

$$F_{2k} = \frac{p\phi P_k}{R + p^2P_k + p(1 - p)S_k} \tag{3.28}$$

$$S_{k+1} = \phi^2 S_k + \Gamma^2 Q \tag{3.29}$$

$$P_{k+1} = (\phi - pF_{2k})P_k\phi + \Gamma^2 Q \tag{3.30}$$

$$P_1 = S_1 = E\{x_1^2\} \tag{3.31}$$

Nahi has shown that P_k as used in the equations above is the variance of $\hat{x}^N_k$.

---

*By the linear estimate we mean that the estimate $\hat{x}^N_k$ is a linear function of the sequence of observations $\mathfrak{z}_k$.

III. 5. 2. Numerical Results

To compare the performance of various solutions of Problem (D-1) we considered the following parameter values:

$\phi_k = -0.8$

$\Gamma_k = 1.0$

$\bar{\omega}_k = 2.0$

$Q_k = 1.0$

$R_k = 1.0$

$p = 0.5$

$\bar{x}_1 = 10.0$

$M_1 = 5.0$

We implemented the LPT solution and the Nahi's solution to this problem on a digital computer. Along with these we also implemented a 'learning with a teacher' type solution in which we considered $\underline{H}_k$ as known and available. One way to compare the performance of these schemes is to compute the mean square error (m. s. e.) of various schemes. We repeated the simulation runs for k = 1 to 25 enough number of times that the sample m. s. e. gave consistent results.* These curves are presented in Figure 3. 2.

---

*Here, to get consistent results we had to repeat the simulation runs 500 times, whereas in Problems (C-1) and (C-2) of Chapter II we had to repeat it only 60 times. When the simulation was repeated with no process noise, we found that 80 runs were enough this time. Therefore it seems that this difference is due to the presence of the process noise $\underline{w}_k$ in Problem D.

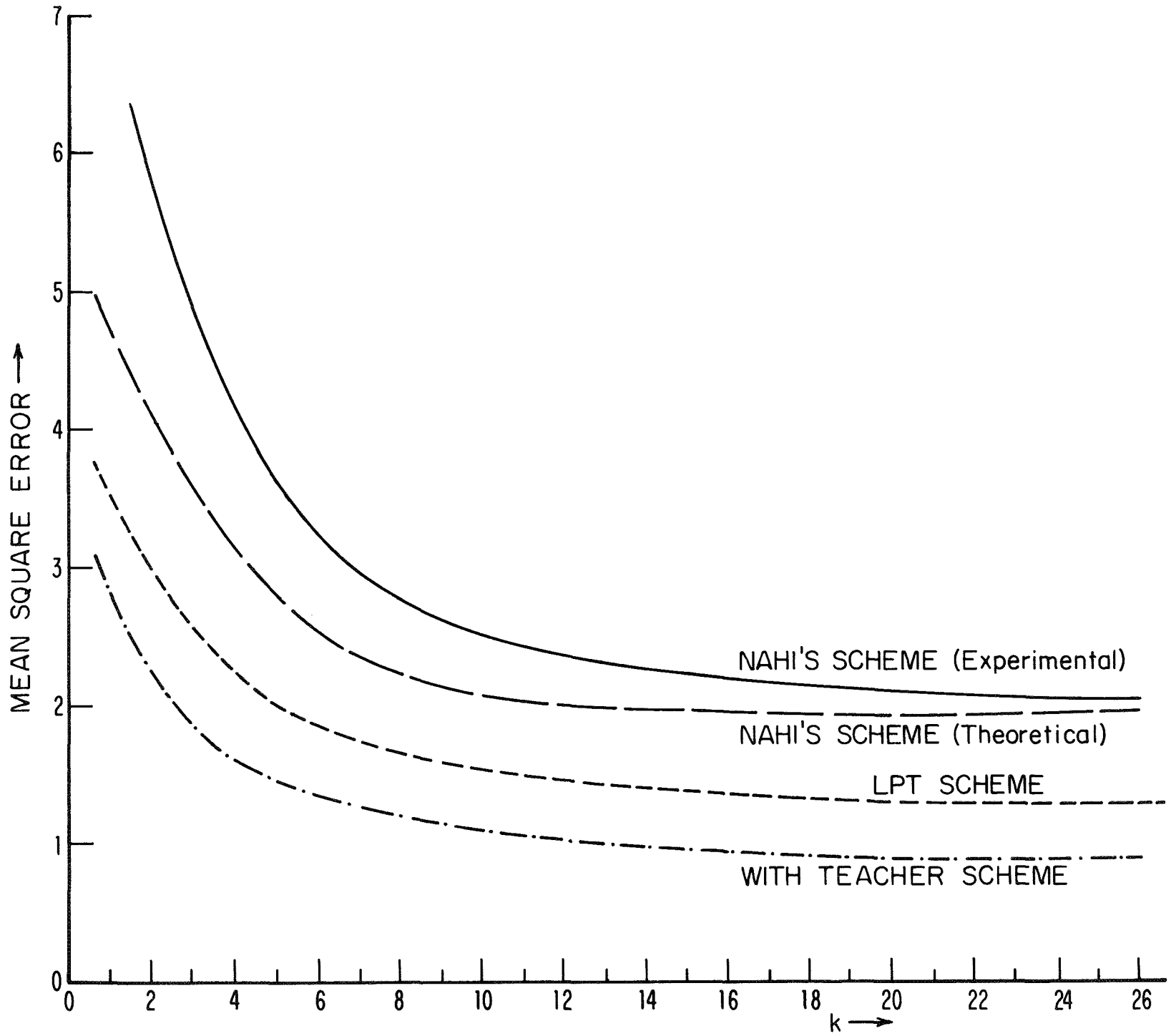


FIG. 3-2 MEAN SQUARE ERROR OF ESTIMATES FOR PROBLEM (D-1).

Figure 3.2 contains two curves of m.s.e. for Nahi's solution. As we noted above P_k of Eq. (3.30) gives the theoretical variance of the estimate $\hat{x}_k^N$. We have plotted this along with the experimental m.s.e. curve. We find that for small k these two curves are far apart. As k increases they come closer but the experimental curve remains above the theoretical curve at all times.

We further note that the mean square error is minimum for the 'learning with a teacher' solution. The m.s.e. of the LPT estimate is approximately twice the m.s.e. of the 'learning with a teacher' estimate. This further confirms our observation of Section II.6 that the average m.s.e. of the LPT estimate is twice the m.s.e. of the 'learning with a teacher' estimate. The m.s.e. of Nahi's estimate, theoretical as well as experimental, is much larger than the m.s.e. for the LPT estimate.

III.6. Summary

In this chapter an estimation problem is considered in which we estimate the state of a Gauss Markov sequence. In addition to the additive Gaussian white noise the observation process for this problem has some multiplicative noise also. This is defined as Problem D.

After showing why the standard techniques will not lead to a feasible solution to this problem, we proceed to formulate a solution using the LPT scheme. The LPT scheme leads to a feasible solution. An example is presented in which the results of this solution are compared with the best linear filter proposed by Nahi [1]. We find that the mean square error of the LPT estimate is always less than that of Nahi's estimate.

References for Chapter III

1. Nahi, N. E., "Optimal Recursive Estimation with Uncertain Observations," IEEE Trans. on Info. Th., Vol. I. T-15, No. 4, pp. 457-462, July 1969.

2. Deutsch, R., 'Estimation Theory,' Prentice Hall, Englewood Cliffs, N. J., 1965.

3. Papulis, A., 'Probability, Random Variables and Stochastic Processes,' McGraw Hill, New York, 1965. Page 189.

4. Bryson, A. E., and Ho, Y. C., 'Applied Optimal Control: Optimization, Estimation and Control,' Blaisdell, Waltham, Massachusetts, 1969.

CHAPTER IV

CONCLUDING REMARKS AND SUGGESTIONS FOR FUTURE WORK

The main contribution of this work is the 'learning with a probabilistic teacher' scheme which can be used to solve a class of unsupervised learning problems. The LPT scheme is formulated in a Bayesian estimation framework. x is considered as an unknown parameter. Some a priori knowledge about it is available as $p_{\underline{x}}(x)$. A sequence of observations $\mathfrak{z}_k = z_1, z_2, z_3, \ldots, z_k$ is given. The Bayesian estimation requires the computation of the posterior density function $p_{\underline{x}/\mathfrak{z}_k}(x; \mathfrak{z}_k)$. In unsupervised learning problems the probability density function for $\underline{z}$ has a mixture form;

$$p_{\underline{z}}(z) = \int_{\mathcal{H}} p_{\underline{z}/\underline{H}}(z; H) p_{\underline{H}}(H) dH$$

When the sequence of the correct classifications $\mathfrak{h}_k = H_1, H_2, H_3, \ldots, H_k$ is given along with the sequence $\mathfrak{z}_k$, the computation of the posterior density function is feasible for a class of problems. For that class, without the knowledge of $\mathfrak{h}_k$, this computation becomes infeasible. In the LPT scheme a label ℓ_k is generated for z_k, which is then treated as the correct value of $\underline{H}_k$. As a result, the computation of the posterior density function now treats the observations as classified samples and hence is feasible.

The generation of the label ℓ_k for the LPT scheme requires computing the probability density function for $\underline{H}_k$, given the observation z_k and the past information. (We have shown that this computation can be carried out easily.) The label ℓ_k is treated as a random variable having

this density function and is generated on the digital computer using a pseudo random number generator. We have shown that the posterior density functions computed using the labels generated this way, converge with probability one to a delta function at the correct value. Also, if an estimate is made from the posterior density function of the LPT scheme, the average variance of such an estimate is twice the variance of an estimate made with the sequence $\mathfrak{h}_k$ known.

The unsupervised learning problems originating in Pattern Recognition context require the estimation of a parameter which has a constant value. The LPT scheme can also be used if the unknown parameter value follows a Gauss Markov sequence.

The posterior density function $p_{\underline{x}/\mathfrak{z}_k}(x;\mathfrak{z}_k)$ may be expressed as

$$p_{\underline{x}/\mathfrak{z}_k}(x;\mathfrak{z}_k) = \int p_{\underline{x}/\mathfrak{z}_k,\underline{\mathfrak{h}}_k}(x;\mathfrak{z}_k,\mathfrak{h}_k)\, p_{\underline{\mathfrak{h}}_k/\mathfrak{z}_k}(\mathfrak{h}_k;\mathfrak{z}_k)\, d\mathfrak{h}_k$$

where $p_{\underline{x}/\mathfrak{z}_k,\underline{\mathfrak{h}}_k}(x;\mathfrak{z}_k,\mathfrak{h}_k)$ is the posterior density function given $\mathfrak{z}_k$ and the sequence of classifications $\mathfrak{h}_k$, and $p_{\underline{\mathfrak{h}}_k/\mathfrak{z}_k}(\mathfrak{h}_k;\mathfrak{z}_k)$ is the probability of occurance of the sequence $\mathfrak{h}_k$ given $\mathfrak{z}_k$. Therefore the computation of $p_{\underline{x}/\mathfrak{z}_k}(x;\mathfrak{z}_k)$ requires computing $p_{\underline{x}/\mathfrak{z}_k,\underline{\mathfrak{h}}_k}$ along all possible $\mathfrak{h}_k$, and algebraically weighting these with the probability of occurance of $\mathfrak{h}_k$ given $\mathfrak{z}_k$. In the LPT scheme we generate a sequence of labels $\mathcal{L}_k$ which has the probability density $p_{\underline{\mathfrak{h}}_k/\mathfrak{z}_k}(\mathcal{L}_k;\mathfrak{z}_k)$. Therefore, while an algebraic weighing is used in computing $p_{\underline{x}/\mathfrak{z}_k}(x;\mathfrak{z}_k)$, the sequence of labels is made random in the LPT scheme such that the expected value of the posterior density $p_{\underline{x}/\mathfrak{z}_k,\underline{\mathcal{L}}_k}$ is $p_{\underline{x}/\mathfrak{z}_k}$. Doing this assures the convergence of the LPT scheme.

The introduction of a randomness to avoid the algebraic weighing, while maintaining the expectation at the correct value, may be considered the central idea of the LPT scheme.

IV.1. Suggestions for Further Work

The present work opens up a number of new problems. Some of these are indicated below.

1. A General Convergence Proof

Consider two estimation procedures A and B. The estimation procedure A gives an estimate $\hat{x}_k^A$ at the kth stage and is assured convergence but does not lead to a feasible solution. Estimation procedure B introduces an extra randomness $\underline{\ell}_k$ in the estimation process and assures that

$$E_{\underline{\ell}_k} [\hat{x}_k^B] = \hat{x}_k^A$$

Under what conditions does the estimate $\hat{x}_k^B$ converge?

In the work presented here the Bayesian estimation is the estimation procedure A and the LPT scheme is the estimation procedure B. We used the martingale theory to prove the convergence in this case. A general convergence theorem may be proved in function theory context.

2. Application of the LPT scheme to Maximum Likelihood Estimation and Stochastic Approximation Framework

The LPT scheme may be formulated in the maximum likelihood and stochastic approximation framework. This formulation may be straightforward but the convergence will have to be established separately.

3. Gauss Markov Sequences with Random ϕ_k, Perfectly Correlated H, etc.

In this work we have indicated how the LPT scheme may be used to estimate the state of a Gauss Markov sequence when h_k is independent white random sequence. The scheme may also be applicable if ϕ is considered as a random variable. Can the LPT scheme be used if $\underline{H}$ is perfectly correlated?

4. Efficient Ways of Generating Random Numbers from a Specified Probability Density Function

The ease of implementation of the LPT scheme depends on the easy generation of a random number from a specified probability density function. We have considered some feasible methods for this purpose. If x_k and z_k are vectors this generation becomes very difficult. Are there any efficient ways of generating random numbers from a specified multi-dimensional probability density function?

ACKNOWLEDGEMENTS

It is difficult to explain in a few words what a great help and a source of encouragement Professor Y. C. Ho has been throughout my career at Harvard. As a faculty and research advisor, all along he has been like a guiding light for me lost in a sea of difficulties. I feel deeply indebted to him.

I am greatful to my colleagues Mr. K. P. S. Prabhu and Mr. J. L. Poage for taking some time off for constructive discussions. I also wish to thank Professor R. E. Kronauer and Professor D. H. Jacobson who read and commented on this manuscript.

Joint Services Distribution List

Asst Director/Research (Rm 3C128)
Office of the Secretary of Defense
Pentagon
Washington, D. C. 20301

Technical Library
DDR and E
Room 3C-122, The Pentagon
Washington, D. C. 20301

Chief, R and D Division (340)
Defense Communications Agency
Washington, D. C. 20305

Director for Materials Sciences
ARPA
Room 3D179, The Pentagon
Washington, D.C. 20301

Major Richard J. Gowen
Tenure Associate Professor
Dept of Electrical Engineering
USAF Academy, Colorado 80840

Defense Documentation Center
Attn: DDC-TCA
Cameron Station
Alexandria, Virginia 22314 (20)

M. A. Rothenberg (STEPD-SD(S))
Scientific Director
Deseret Test Center
Bldg 100, Soldiers' Circle
Fort Douglas, Utah 84113

Mr. H. E. Webb, Jr (EMBIS)
Rome Air Development Center
Griffiss Air Force Base, New York 13440

Central Intelligence Agency
Attn: CRS/ADD/PUBLICATIONS
Washington, D.C. 20505

Hq. USAF (AFRDDD)
The Pentagon
Washington, D.C. 20330

Hq. USAF (AFRDDG)
The Pentagon
Washington, D.C. 20330

Hq. USAF (AFRDSD)
The Pentagon
Washington, D.C. 20330
Attn: LTC C. M. Waespy

Colonel E. P. Gaines, Jr.
ACDA/FO
1901 Pennsylvania Avenue N. W.
Washington, D.C. 20451

Lt. Col. H. W. Jackson (SREE)
Chief, Electronics Division
Directorate of Engineering Sciences
Air Force Office of Scientific Research
Arlington, Virginia 22209 (5)

Dr. I. R. Mirman
Hq AFSC (SCGP)
Andrews Air Force Base
Washington, D. C. 20331

Commanding General
USACDC Institute of Land Combat
Attn: Technical Library, Rm 636
2461 Eisenhower Avenue
Alexandria, Virginia 22314

Rome Air Development Center
Attn: Documents Library (EMTLD)
Griffiss Air Force Base
New York 13440

MIT Lincoln Laboratory
Attn: Library A-082
P. O. Box 73
Lexington, Mass. 02173

Dr. L. M. Hollingsworth
AFCRL (CRN)
L. G. Hanscom Field
Bedford, Massachusetts 01730

VELA Seismological Center
300 North Washington Street
Alexandria, Virginia 22314

Hq. ESD (ESTI)
L. G. Hanscom Field
Bedford, Massachusetts 01730 (2)

Prof. R. H. Rediker
Electrical Engineering, Professor
MIT
Building 13-3050
Cambridge, Massachusetts 02139

AFAL (AVT) Dr. H. V. Noble
Electronics Technology Division
Air Force Avionics Laboratory
Wright-Patterson AFB, Ohio 45433

Director
Air Force Avionics Laboratory
Wright-Patterson AFB
Ohio 45433

AFAL (AVTA/R. D. Larson)
Wright-Patterson AFB
Ohio 45433

Director of Faculty Research
Dept. of the Air Force
U. S. Air Force Academy
Colorado Springs, Colorado 80840

Academy Library (DFSLB)
USAF Academy
Colorado Springs, Colorado 80840

Director
Aerospace Mechanics Sciences
Frank J. Seiler Research Lab (OAR)
USAF Academy
Colorado Springs, Colorado 80840

Director, USAF PROJECT RAND
Via: Air Force Liaison Office
The RAND Corporation
Attn: Library D
1700 Main Street
Santa Monica, California 90406

HQ SAMSO (SMTAE/Lt. Belate)
AF Unit Post Office
Los Angeles, California 90045

Miss R. Joyce Harman
Project MAC, Room 810
545 Main Street
Cambridge, Mass. 02139

AUL3T-9663
Maxwell AFB, Alabama 36112

AFETR Technical Library
(ETV, MU-135)
Patrick AFB, Florida 32925

ADTC (ADBPS-12)
Eglin AFB, Florida 32542

Mr. B. R. Locke
Technical Adviser, Requirements
USAF Security Service
Kelly Air Force Base, Texas 78241

Hq. AMD (AMR)
Brooks AFB, Texas 78235

USAFSAM (SMKOR)
Brooks AFB, Texas 78235

Commanding General
Attn: STEWS-RE-L, Technical Library
White Sands Missile Range
New Mexico 88002 (2)

Hq. AEDC (AETS)
Arnold AFS, Tennessee 37389

USAF
European Office of Aerospace Research
APO, New York 09667

Director
Physical and Engineering Sciences Div.
3045 Columbia Pike
Arlington, Virginia 22204

Commanding General
U. S. Army Security Agency
Attn: IARD-T
Arlington Hall Station
Arlington, Virginia 22212

Commanding General
U. S. Army Material Command
Attn: AMCRD-TP
Washington, D.C. 20315

Commanding Officer
Harry Diamond Laboratories
Attn: Dr. Berthold Altman (AMXDO-TI)
Connecticut Avenue & Van Ness St. N. W.
Washington, D. C. 20438

Chief
Missile Electronic Warfare Tech Area
(AMSEL-WL-M)
U. S. Army Electronics Command
White Sands Missile Range
New Mexico 88002

Commanding Officer (AMXRD-BAT)
U. S. Army Ballistics Research Lab.
Aberdeen Proving Ground
Aberdeen, Maryland 21005

Technical Director
U. S. Army Limited War Laboratory
Aberdeen Proving Ground
Aberdeen, Maryland 21005

Commanding Officer
U. S. Army Engineer Topographic Labs.
Attn: STINFO Center
Fort Belvoir, Virginia 22060

U. S. Army Munitions Command
Attn: Science & Technology Info. Branch
Building 59
Picatinny Arsenal, SMUPA-RT-S
Dover, New Jersey 07801

U. S. Army Mobility Equipment Research
and Development Center
Attn: Technical Document Center
Building 315
Fort Belvoir, Virginia 22060

Commanding Officer (AMSEL-BL-WS-R
Atmospheric Sciences Laboratory
U. S. Army Electronics Command
White Sands Missile Range
New Mexico 88002

Dr. Herman Robl
Deputy Chief Scientist
U. S. Army Research Office (Durham)
Box CM, Duke Station
Durham, North Carolina 27706

Richard O. Ulsh (CRDARD-IP)
U. S. Army Research Office (Durham)
Box CM, Duke Station
Durham, North Carolina 27706

Technical Director
(SMUFA-A3000-107-1)
Frankford Arsenal
Philadelphia, Pa. 19137

Redstone Scientific Info. Center
Attn: Chief, Document Section
U. S. Army Missile Command
Redstone Arsenal, Alabama 35809

Commanding General
U. S. Army Missile Command
Attn: AMSMI-RR
Redstone Arsenal, Alabama 35809

Commanding General
U. S. Army Strategic Comm. Command
Attn: SCC-CG-SAE
Fort Huachuca, Arizona 85613

Commanding Officer
Army Materials and Mechanics
Research Center
Attn: Dr. H. Priest
Watertown Arsenal
Watertown, Mass. 02172

Commandant
U. S. Army Air Defense School
Attn: Missile Science Div., C&S Dept.
P. O. Box 9390
Fort Bliss, Texas 79916

Commandant
U. S. Army Command and General
Staff College
Attn: Acquisitions, Lib. Div.
Fort Leavenworth, Kansas 66027

Mr. Norman J. Field, AMSEL-RD-S
Chief, Office of Science & Technology
Research & Development Directorate
U. S. Army Electronics Command
Fort Monmouth, New Jersey 07703

Mr. Robert O. Parker, AMSEL-RD-S
Executive Secretary, TAC/JSEP
U. S. Army Electronics Command
Fort Monmouth, New Jersey 07703

Commanding General
U. S. Army Electronics Command
Fort Monmouth, N. J. Jersey 07703
Attn: AMSEL-SC
DL
GG-DD
XL-D
XL-DT
BL-FM-P
CT-D
CT-R
CT-S
CT-L (Dr. W. S. McAfee)
CT-O
CT-I
CT-A
NL-D (Dr. H. Bennett)
NL-A
NL-C
NL-P
NL-P-2
NL-R
NL-S
KL-D
KL-I
KL-E
KL-S
KL-SM
KL-T
VL-D
VL-F
WL-D

Dr. Alvin D. Schnitzler
Institute for Defense Analyses
Science and Technology Division
400 Army-Navy Drive
Arlington, Virginia 22202

Director (NV-D)
Night Vision Laboratory, USAECOM
Fort Belvoir, Virginia 22060

Commanding Officer
Atmospheric Sciences Laboratory
U. S. Army Electronics Command
White Sands Missile Range
New Mexico 88002

Code 8050
Maury Center Library
Naval Research Laboratory
Washington, D. C. 20390

Dr. A. G. Jordan
Head of Dept. of Electrical Engineering
Carnegie-Mellon University
Pittsburgh, Pennsylvania 15213

Project Manager
NAVCON
Attn: Harold H. Bahr (AMCPM-NS-TM)
Building 439
U. S. Army Electronics Command
Fort Monmouth, New Jersey 07703

Director, Electronic Programs
Attn: Code 427
Dept. of the Navy
Washington, D.C. 20360 (3)

Commander
Naval Security Group Command
Naval Security Group Headquarters
Attn: G43
3801 Nebraska Avenue
Washington, D.C. 20390

Director
Naval Research Laboratory
Washington, D.C. 20390
Attn: Code 2027 (6)
Dr. W. C. Hall, Code 7000 (1)
Dr. A. Brodzinsky (1)
Supt. Elec. Div.

Dr. G. M. R. Winkler
Director, Time Service Division
U. S. Naval Observatory
Washington, D.C. 20390

Naval Air Systems Command
AIR 03
Washington, D.C. 20360 (2)

Naval Ship Systems Command
SHIP 031
Washington, D.C. 20360

Naval Ship Systems Command
SHIP 035
Washington, D.C. 20360

U. S. Naval Weapons Laboratory
Dahlgren, Virginia 22448

Naval Electronic Systems Command
ELEX 03, Rm 2534 Main Navy Bldg.
Dept. of the Navy
Washington, D.C. 20360 (2)

Government Documents Dept.
University of Iowa Libraries
Iowa City, Iowa 52240

Commander
U. S. Naval Ordnance Laboratory
Attn: Librarian
White Oak, Maryland 21502 (2)

Director
Naval Research Laboratory
Attn: Library, Code 2029 (ONRL)
Washington, D.C. 20390 [2]

Hollander Associates
P. O. Box 2276
Fullerton, California 92633

Illinois Institute of Technology
Dept. of Electrical Engineering
Chicago, Illinois 60616

The University of Arizona
Dept. of Electrical Engineering
Tucson, Arizona 85721

Utah State University
Dept. of Electrical Engineering
Logan, Utah 84321

Case Institute of Technology
Engineering Division
University Circle
Cleveland, Ohio 44106

Carl E. Baum, Capt.
AFWL (WLRE)
Kirtland AFB, New Mexico 87117

Lenkurt Electric Co., Inc.
1105 County Road
San Carlos, California 94070
Attn: Mr. E. K. Peterson

Dr. F. R. Charvat
Union Carbide Corporation
Materials Systems Division
Crystal Products Dept.
8888 Balboa Avenue
P. O. Box 23017
San Diego, California 92123

Director
U. S. Army Advanced Materiel
Concepts Agency
Washington, D. C. 20315

Electromagnetic Compatibility
Analysis Center
(ECAC), Attn: ACOAT
North Severn
Annapolis, Maryland 21402

Dept. of Electrical Engineering
Rice University
Houston, Texas 77001

Research Laboratories for the
Engineering Sciences
School of Engineering and Applied
Science
University of Virginia
Charlottesville, Virginia 22903

Dept. of Electrical Engineering
Clippinger Laboratory
Ohio University
Athens, Ohio 45701

Lehigh University
Dept. of Electrical Engineering
Bethlehem, Pennsylvania 18015

Professor James A. Cadzow
Dept. of Electrical Engineering
State Univ. of New York at Buffalo
Buffalo, New York 14214

Director
Office of Naval Research Branch Office
495 Summer Street
Boston, Mass. 02210

Commander (AOL)
Naval Air Development Center
Johnsville, Warminster, Pa. 18974
Attn: NADC Library

Commander (Code 753)
Naval Weapons Center
Attn: Technical Library
China Lake, California 93555

Commanding Officer
Naval Weapons Center
Corona Annex
Attn: Library
Corona, California 91720

Commanding Officer (56322)
U. S. Naval Missile Center
Point Mugu, California 93041

W. A. Eberspacher, Assoc. Head
Systems Integration Division
Code 5340A
U. S. Naval Missile Center
Point Mugu, California 93041

Commander
Naval Electronics Laboratory Center
Attn: Library
San Diego, California 92152 (2)

Deputy Director and Chief Scientist
Office of Naval Research Branch Office
1030 East Green Street
Pasadena, California 91101

Library (Code 2124)
Technical Report Section
Naval Postgraduate School
Monterey, California 93940

Glen A. Myers (Code 52 Mv)
Assoc. Prof. of Electrical Engineering
Naval Postgraduate School
Monterey, California 93940

Commanding Officer (Code 2064)
Navy Underwater Sound Laboratory
Fort Trumbull
New London, Conn. 06320

Commanding Officer
Naval Avionics Facility
Indianapolis, Indiana 46241

Dr. H. Harrison, Code RRE
Chief, Electrophysics Branch
National Aeronautics and Space Admn.
Washington, D.C. 20546

NASA Lewis Research Center
Attn: Library
21000 Brookpark Road
Cleveland, Ohio 44135

Los Alamos Scientific Laboratory
Attn: Reports Library
P. O. Box 1663
Los Alamos, New Mexico 87544

Mr. M. Zane Thornton, Chief
Network Engineering, Communications
and Operations Branch
Lister Hill National Center for
Biomedical Communications
8600 Rockville Pike
Bethesda, Maryland 20014

U. S. Post Office Dept.
Library - Room 6012
12th & Pennsylvania Ave, N. W.
Washington, D.C. 20260

Director
Research Lab of Electronics
MIT
Cambridge, Mass. 02139

Mr. Jerome Fox
Research Coordinator
Polytechnic Institute of Brooklyn
333 Jay Street
Brooklyn, New York 11201

Director
Columbia Radiation Laboratory
Columbia University
538 West 120th Street
New York, New York 10027

Director
Coordinated Science Laboratory
University of Illinois
Urbana, Illinois 61801

Director
Stanford Electronics Laboratories
Stanford University
Stanford, California 94305

Director
Microwave Physics Laboratory
Stanford University
Stanford, California 94305

Director
Electronics Research Laboratory
University of California
Berkeley, California 94720

Director
Electronic Sciences Laboratory
University of Southern California
Los Angeles, California 90007

Director
Electronics Research Center
The University of Texas at Austin
Engineering-Science Bldg 110
Austin, Texas 78712

Division of Engineering and Applied Physics
130 Pierce Hall
Harvard University
Cambridge, Mass. 02138

Dr. G. J. Murphy
The Technological Institute
Northwestern University
Evanston, Illinois 60201

Dr. John C. Hancock, Head
School of Electrical Engineering
Purdue University
Lafayette, Indiana 47907

Dept. of Electrical Engineering
Texas Technological College
Lubbock, Texas 79409

Aerospace Corporation
P. O. Box 95085
Los Angeles, California 90045
Attn: Library Acquisitions Group

Professor Nicholas George
California Institute of Technology
Pasadena, California 91109

Aeronautics Library
Graduate Aeronautical Laboratories
California Institute of Technology
1201 E. California Blvd
Pasadena, California 91109

The Johns Hopkins University
Applied Physics Laboratory
Attn: Document Librarian
8621 Georgia Avenue
Silver Spring, Maryland 20910

Hunt Library
Carnegie-Mellon University
Schenley Park
Pittsburgh, Pa. 15213

Dr. Leo Young
Stanford Research Institute
Menlo Park, California 94025

Chairman, Electrical Engineering
Arizona State University
Tempe, Arizona 85281

Engineering & Mathematical
Sciences Library
University of Calif. at L. A.
405 Hilgred Avenue
Los Angeles, California 90024

Sciences-Engineering Library
University of California
Santa Barbara, California 93106

Prof. Joseph E. Rowe
Chairman, Dept. of Electrical Engin.
The University of Michigan
Ann Arbor, Michigan 48104

Dr. W. R. LePage, Chairman
Syracuse University
Dept. of Electrical Engineering
Syracuse, New York 13210

Yale University
Dept. of Engineering and Applied Science
New Haven, Conn. 06520

Airborne Instruments Laboratory
Deerpark, New York 11729

Raytheon Company
Research Division Library
28 Seyon Street
Waltham, Massachusetts 02154

Dr. Sheldon J. Welles
Electronic Properties Information Center
Mail Station E-175
Hughes Aircraft Company
Culver City, California 90230

Dr. Robert E. Fontana
Dept. of Electrical Engineering
Air Force Institute of Technology
Wright-Patterson AFB, Ohio 45433

Dr. John R. Ragazzini, Dean
School of Engineering and Science
New York University
University Heights
Bronx, New York 10453

Sylvania Electronic Systems
Applied Research Laboratory
Attn: Documents Librarian
40 Sylvan Road
Waltham, Mass. 02154

Unclassified
Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author)
Division of Engineering and Applied Physics
Harvard University
Cambridge, Mass. 02138

2a. REPORT SECURITY CLASSIFICATION
Unclassified

2b. GROUP

3. REPORT TITLE
LEARNING WITH A PROBABILISTIC TEACHER

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)
Interim technical report

5. AUTHOR(S) (First name, middle initial, last name)
Ashok K. Agrawala

6. REPORT DATE
May 1970

7a. TOTAL NO. OF PAGES
120

7b. NO. OF REFS
31

8a. CONTRACT OR GRANT NO.
N00014-67-A-0298-0006 and NASA Grant NGL 22-007-143

b. PROJECT NO.

c.

d.

9a. ORIGINATOR'S REPORT NUMBER(S)
Technical Report No. 611

9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)

10. DISTRIBUTION STATEMENT
This document has been approved for public release and sale; its distribution is unlimited. Reproduction in whole or in part is permitted by the U. S. Government.

11. SUPPLEMENTARY NOTES

12. SPONSORING MILITARY ACTIVITY
Office of Naval Research

13. ABSTRACT

Estimation or learning problems arise in practical systems in many ways. Depending on the learning information available, the estimation problem may be supervised or unsupervised. Bayesian estimation may be used for both these problems. The Bayesian solution of a supervised learning problem is reasonably simple while the unsupervised Bayesian learning is enormously complex. A practical way of solving an unsupervised learning problem is to convert it into a supervised learning problem by labelling the observation before using it for learning. Decision directed learning scheme uses the result of a decision process as the label. The computations for this scheme are feasible but the resulting estimates do not converge to the correct value.

A learning scheme, 'learning with a probabilistic teacher', is proposed in which a label is generated as a random variable from an appropriate probability density function. This scheme leads to a feasible solution to an unsupervised learning problem and assures the convergence of the estimate to the correct value. The average mean square error of the resulting estimate is twice the mean square error of the 'learning with a teacher' estimate. This learning scheme can also be used to estimate the state of a Gauss Markov sequence when the observation process has additive as well as multiplicative noise.

DD FORM 1 NOV 65 1473 (PAGE 1)
S/N 0101-807-6811

Unclassified
Security Classification

A-31408

Unclassified
Security Classification

14. KEY WORDS	LINK A ROLE	LINK A WT	LINK B ROLE	LINK B WT	LINK C ROLE	LINK C WT
Unsupervised Learning						
Estimation						
Learning						
Pattern Classification						
Bayesian Estimation						
Bayesian Learning						
Learning with/without teacher						

DD FORM 1 NOV 65 1473 (BACK)
S/N 0101-807-6801

Unclassified
Security Classification

A-31409