LEARNING AUTOMATA POSSESSING ERGODICITY OF THE MEAN : THE TWO ACTION CASE

M. A. L. Thathachar* and B. J. Oommen**

SCS-TR-24

May 1983

* Department of Electrical Engineering, Indian Institute of Science, Bangalore, INDIA: 560 012

** School of Computer Science, Carleton University, Ottawa, K1S 5B6 CANADA

To be published in the IEEE Trans. on Systems, Man and Cybernetics. Details of the Issue of Publication are not known.

LEARNING AUTOMATA POSSESSING ERGODICITY OF THE MEAN : THE TWO ACTION CASE

M. A. L Thathachar[+] and B. J. Oommen[++]

ABSTRACT

Learning automata which update their action probabilities on the basis of the responses they get from an environment are considered in this paper. The automata update the probabilities whether the environment responds with a reward or a penalty. An automaton is said to possess Ergodicity of the Mean (EM) if the mean action probability is the total state probability of an ergodic Markov chain. The only known algorithm which is Ergodic in the Mean (EM) is the Linear Reward-Penalty (L_{RP}) scheme. For the 2-action case necessary and sufficient conditions have been derived for nonlinear updating schemes to be Ergodic in the Mean (EM). The method of controlling the rate of convergence of this scheme has been presented. In particular a generalized linear algorithm has been proposed which is superior to the Linear Reward-Penalty (L_{RP}) scheme. The expression for the variance of the limiting action probabilities of this scheme has been derived. The technique of designing the optimal linear automaton in this family has also been considered. Methods to decrease the variance for the general nonlinear scheme have been discussed.

It has been shown that the set of absolutely expedient schemes and the set of schemes which possess ergodicity of the mean are mutually disjoint.

[+] Department of Electrical Engineering, Indian Institute of Science, Bangalore : 560012, India.
[++] Department of Computer Science, Carleton University, Ottawa, Ontario : K1S 5B6, Canada

Index Terms : Ergodic Learning Automata, Ergodicity of the Mean, Linear and Nonlinear Learning Schmes, Variance Reduction Techniques.

I. INTRODUCTION

Learning automata which interact with an environment have been used to model biological learning systems and have a variety of applications such as parameter optimization, statistical decision making, adaptive control of systems and the routing of telephone calls[6]. The environment offers the automaton a finite set of actions, and it is constrained to choose one of them. Depending on the choice of the automaton the environment either penalizes or rewards it. A learning automaton is one which learns the "most favourable" action as time proceeds and which chooses this action "more frequently" in some sense as it learns.

Learning automata can be broadly classified into three families : deterministic, Fixed Structure Stochastic (FSS) and Variable Structure Stochastic (VSS). Deterministic learning automata, such as the Tsetlin and the Krinsky[1,2] automata, are those in which both the transition matrix and the output matrix are deterministic. Fixed Structure Stochastic (FSS) automata are those which have time invariant stochastic matrices determining their transitions; for, with no loss of generality the output matrix can be assumed to be deterministic[3]. The Krylov automaton is one such automaton[2].

Whereas deterministic and FSS automata are easy to implement they have the disadvantage that if they choose a particular action at a certain time, they will choose the same action at the next time instant irrespective of the response of the automaton except when they are in their end states. This is not desirable especially when the number of states is large or the penalty probabilities associated with the actions are close to each

other. To overcome this drawback the notion of a Variable Structure Stochastic (VSS) automaton was introduced [4]. Automata of this type have stochastic transition matrices whose elements are updated as the learning process evolves. It is interesting to note that VSS automata can be constructed by merely formulating a scheme by which the action probabilities can be updated.

Action probability updating schemes studied in the literature fall into two major classes -- those which have absorbing barriers, and those which are ergodic. Whereas in the former class the value of the limiting probability vector depends on the initial action probabilities, in the latter the distribution of the limiting probabilities are independent of the distribution of the initial action probabilities. The latter is a desirable feature especially since automata of this type do not get "locked" into any one action -- which is particularly undesirable when the penalty probabilities are time varying -- that is, the environment is nonstationary.

The simplest ergodic scheme known is probably the Linear Reward-Penalty (L_{RP}) scheme. In this case the action probability decrements are made linearly proportional to the probabilities themselves and are made for reward as well as penalty responses of the environment. The limiting probability vector converges in distribution, and the parameters of this distribution have been known for the symmetric version of the L_{RP} scheme which is a one parameter probability updating algorithm[9]. Ergodic schemes using the concept of stochastic approximation have been proposed and investigated by Lakshmivarahan[11], Flerov[12], Tsypkin and

Poznyak[13,14] and El Fattah[15,16].

To help introduce the contributions of this paper we need the following definition.

Definition I : A learning scheme is said to be Ergodic in the Mean (EM) or equivalently possess Ergodicity of the Mean (EM) if the mean action probability is the state probability* of an ergodic Markov chain.

Remark : The concept of ergodicity of the mean appears important to us because it is one of the simple ways in which the mean of the action probability vector can be made to possess certain desirable characterstics. All the well studied properties of an ergodic Markov chain such as limiting distribution, rate of convergence etc. can be now caried over to the learning automaton possessing this property. The literature reports of only one scheme that is known to be EM and this is the symmetric L_{RP} scheme. We thus refer to the latter as the L_{EM} scheme.

In this paper we consider the general problem of the two action probability updating scheme possessing ergodicity of the mean. The updating algorithm is given in terms of two nonlinear functions $\emptyset(.)$ and $\theta(.)$. Two conditions involving these functions necessary and sufficient for EM have been derived. Whereas the first of these conditions resembles the one proven to be necessary and sufficient for absolute expediency [5,6,8], the second is a linear constraint involving the functions and a constant. The latter constant is the only parameter which

* Also called "absolute probability" or "unconditional probability".

controls the rate of convergence of the scheme. Further, the other parameters in the scheme can be used to control the variance of the limiting action probabilities. The process of designing a nonlinear EM automaton superior to the correspnding L_{RP} automaton has also been suggested.

In particular we have studied a Generalized Linear scheme which is EM. The latter is referred to as the GL_{EM} scheme. This scheme has two parameters and it differs from the symmetric L_{RP} scheme in that the probability decrement is a linear function of the probabilities themselves and is not merely proportional to them. Expressions for the mean and the variance of the limiting vector have been obtained. This scheme has one parameter more than the L_{RP} scheme, and this new parameter solely controls the rate of convergence of the probabilities. Using the two parameters the variance of the limiting probabilities can be minimized for a desirable rate of convergence.

The organization of the paper is as follows. We first introduce the terminology used in the literature and explain the Linear Reward-Penalty (L_{RP}) automaton. We then present the conditions for the general nonlinear updating algorithm to be EM. The Generalized Linear EM (GL_{EM}) scheme is then studied and its variance and convergence properties investigated. We finally present simulation results which demonstrate the learning capabilities of the automata discussed.

I.1 Fundamentals

The automaton selects an action a(n) at a time instant 'n'. a(n) is any one of a finite set $\{a_1, \ldots, a_R\}$ and is selected on

the basis of a R X 1 probability vector $\underline{p}(n)$ where :

$$p_i(n) = Pr [a(n) = a_i] \quad \text{with} \quad \sum_{i=1}^{R} p_i(n) = 1$$

The selected action interacts with a random environment which gives out a response b(n) at the same time instant. b(n) is either 0 or 1, the latter being called the penalty. The quantity c_i is defined below is referred to as the penalty probability.

$$c_i = Pr [b(n) = 1 \mid a(n) = a_i] \qquad (i = 1 \text{ to } R)$$

Thus the environment is characterized by the set of penalty probabilities. The automaton updates the vector $\underline{p}(n)$ on the basis of b(n) and then a new action is chosen at (n+1).

The $\{c_i\}$ are unknown initially and it is desired that as a result of the feedback received from the environment, the automaton will ultimately choose the action with the minimum c_i more frequently in the expected sense.

The average penalty received at the nth time instant is

$$M(n) = \sum_{i=1}^{R} p_i(n) c_i$$

With no apriori information the automaton chooses the actions with equal probability. The expected penalty is thus initially M_o,

$$M_o = \sum_{i=1}^{R} p_i(0) c_i = \frac{1}{R} \sum_{i=1}^{R} c_i \quad (\text{since } p_i(0) = 1/R)$$

An automaton is said to learn <u>expediently</u> if, as time tends towards infinity, the expected penalty is less than M_o.

The automaton is <u>absolutely expedient</u> if

$$E [M(n+1) \mid p(n)] < M(n)$$

Note that in this case M(n) is a supermartingale [8].

Throughout this paper we shall be considering 2-action automata, i.e., with R=2. The properties of the general R-action

EM scheme are currently being investigated.

I.2 The L_{EM} Scheme

The Linear Reward-Penalty (L_{RP}) scheme which is a probability updating algorithm having two parameters a,b < 1 is given below.

$$\begin{aligned} p_i(n+1) &= a\, p_i(n) && \text{if } a(n) = a_i \text{ and } b(n) = 1 \\ &= b p_i(n) && \text{if } a(n) = a_j \text{ and } b(n) = 0 \\ &= (1-b) + b\, p_i(n) && \text{if } a(n) = a_i \text{ and } b(n) = 0 \\ &= (1-a) + a\, p_i(n) && \text{if } a(n) = a_j \text{ and } b(n) = 1 \end{aligned}$$

To simplify the notation, unless explicitly stated we use p_i to refer to the probability $p_i(n)$. In this form of the L_{RP} scheme $E[p_i(n+1)|p_i]$ has the expression :

$$E[p_i(n+1)|p_i] = p_i^2\,(a-b)(c_i-c_j) + p_i\{1-c_i(1-b)-c_j(1+b-2a)\}+c_j(1-a)$$

where i,j=1,2 and $i \neq j$.

Observe that $E[p_i^k(n+1)]$ is dependent on $E[p_i^r(n+1)]$ for some $r \geq k$ for all $k > 1$. Because of this the form of the limiting distribution of the general L_{RP} scheme is unknown. However, if b=a, the term containing p_i^2 disappears from the above expression and renders it EM. Using distance diminshing operators[9,10] it can be shown that when b=a the limiting distribution of p_i has the following mean and variance:

$$E[\, p_i(\infty)\,] = c_j / (c_i + c_j) \qquad i,j = 1,2\ ;\ i \neq j$$

$$Var[\, p_i(\infty)\,] = \frac{c_1\, c_2\ (1-a)}{(c_1 + c_2)^2\ [(1-a) + 2a(c_1 + c_2)]}$$

We refer to the symmetric L_{RP} scheme with b=a as the L_{EM} scheme (for Linear scheme possessing Ergodicity of the Mean). This is the only known probability updating algorithm which is

EM. It is expedient. Further, the parameter 'a' controls both the rate of convergence and the variance of the limiting distribution, both of which increase with 'a'. In a later section we shall present a Generalized Linear EM (GL_{EM}) scheme that has two parameters and which provides greater flexibility with regard to controlling the rate of convergence and the limiting variance.

II NONLINEAR SCHEMES ERGODIC IN THE MEAN

We shall first consider the general problem of designing nonlinear EM learning schemes. Two conditions necessary and sufficient for probability updating schemes to be EM have been derived. The conditions involve two arbitary functions $\emptyset(.)$ and $\theta(.)$. The first of these conditions is similar to the conditions required to guarantee absolute expediency [5,6,8]. The second condition constrains the two arbitary functions introduced to be linearly dependent.

The general form of the nonlinear probability updating algorithm is given below. As in the previous section, i,j=1,2 with $i \neq j$. Throughout this section except when explicitly stated, p_i refers to $p_i(n)$.

$$\begin{aligned} p_i(n+1) &= \emptyset(p_i) && \text{if } a(n) = a_i \text{ and } b(n) = 1 \\ &= \theta(p_i) && \text{if } a(n) = a_j \text{ and } b(n) = 0 \\ &= 1 - \emptyset(p_j) && \text{if } a(n) = a_j \text{ and } b(n) = 1 \\ &= 1 - \theta(p_j) && \text{if } a(n) = a_i \text{ and } b(n) = 0 \end{aligned} \qquad (1)$$

The above equations specify that if the automaton chose action a_i and it was penalized, then the probability p_i is updated to $\emptyset(p_i)$. Further, if it selected action a_j and was rewarded then the probability p_i is updated to $\theta(p_i)$, thus changing p_j from its current value to $1-\theta(p_i)$. Note that the scheme is symmetric with

respect to the actions, i.e., if the scheme updated p_i in any circumstance which involved p_i (p_j), it should update p_j identically when the situation involved p_j (p_i). To ensure order that $p_i(n+1)$ and $p_j(n+1)$ remain to be proabilities and that the scheme is of a reward-penalty nature the most obvious constraint on $\emptyset(.)$ and $\theta(.)$ is :

$$1 \geq p_i \geq \emptyset(p_i), \quad \theta(p_i) \geq 0, \quad \theta(0) = 0$$

The conditions for ergodicity for the above scheme to be EM are now derived.

Theorem I

The Necessary and Sufficient conditions for the nonlinear updating algorithm scheme defined by (1) to be EM in all stationary random environments are :

$$\theta(p_i) / p_i = \theta(p_j) / p_j = w(p_i, p_j)$$
$$\emptyset(p_i) + \theta(p_j) = d \qquad (2)$$

where, $i,j=1,2$ and $i \neq j$, and $d < 1$.

Proof : Since $\underline{p}(n+1)$ has the following distribution :

$$\begin{aligned} p_i(n+1) &= \emptyset(p_i) && \text{with prob. } p_i c_i \\ &= \theta(p_i) && \text{with. prob. } p_j (1-c_j) \\ &= 1 - \emptyset(p_j) && \text{with prob. } p_j c_j \\ &= 1 - \theta(p_j) && \text{with prob. } p_i (1-c_i) \end{aligned}$$

$$E[p_i(n+1) | p_i] = p_i c_i \{ \emptyset(p_i) + \theta(p_j) \} - p_j c_j \{ \emptyset(p_j) + \theta(p_i) \} + \{ p_i - p_i c_i + c_j - c_j p_i \} + \{ p_j \theta(p_i) - p_i \theta(p_j) \} \qquad (3)$$

On taking expectations again we observe that if $E[\underline{p}(n+1)]$ is to be the state probability vector of a Markov chain the right hand side of the above equation must be a linear function in p_i and p_j. Since the first two terms involve c_i and c_j (which could be arbitary) and the nonlinear functions $\emptyset(.)$ and $\theta(.)$ the

sufficient and necessary condition for these terms to be linear in p_i and p_j is $\emptyset(p_i)+\theta(p_j)=d$, where $i,j=1,2$ with $i \neq j$.

We investigate the conditions on $\emptyset(.)$ and $\theta(.)$ for the last term in (3) to be of the form

$$p_j\,\theta(p_i) - p_i\,\theta(p_j) = ap_i + bp_j \qquad (4)$$

Since $p_i+p_j=1$ any constant added to the right hand side of (4) can be absorbed into the coefficients of p_i and p_j. Thus this is the most general form of a linear function in p_i and p_j.

The boundary conditions $p_i=0$ and $p_j=0$ require that the coefficients a and b must be equal to $\pm\,\theta(0)$ which is 0. Hence,

$$\theta(p_i) / p_i = \theta(p_j) / p_j$$

$p_i=0 \Rightarrow -p_i\,\theta(0) = bp_j \Rightarrow b=\theta(0)=0$
$p_j=0 \Rightarrow -p_j\,\theta(0) = ap_i \Rightarrow a=-\theta(0)=0$
$\therefore p_j\theta(p_i) - p_i\theta(p_j)=0 \Rightarrow \frac{\theta(p_i)}{p_i} = \frac{\theta(p_j)}{p_j}$

is a necessary and sufficient condition for the scheme to be EM. Hence the theorem.

Theorem II

Nonlinear EM schemes defined by (1) are always expedient.

Proof : Substituting the necessary and sufficient conditions for a nonlinear scheme to be EM into (3), we get,

$$E[p_i(n+1) \mid p_i] = p_i\,\{\,1 - (c_i+c_j)(1-d)\,\} + (1-d)\,c_j(p_i+p_j)$$
$$= p_i\,\{\,1 - c_i(1-d)\} + (1-d)c_j p_j$$

Since $p_i + p_j = 1$, we multiply the last term by p_i+p_j. Taking expectations again yields,

$$E[p_i(n+1)] = E[p_i(n)]\,\{\,1 - c_i(1-d)\,\} + E[p_j(n)]\,\{(1-d)\,c_j\}$$

Thus $E[\underline{p}(n+1)] = Q^T\,E[\underline{p}(n)]$, where,

$$Q = \begin{bmatrix} 1-c_1(1-d) & c_1(1-d) \\ c_2(1-d) & 1-c_2(1-d) \end{bmatrix}$$

The above defines an ergodic Markov process if $d < 1$ [7]. The limiting value of $E[p_i(\infty)]$ is obtained by solving

$$E[\underline{p}(\infty)] = Q^T\,E[\underline{p}(\infty)]$$

same as L_{RP}

This final value of $E[p_i(\infty)]$ is $c_j / (c_i + c_j)$. Thus every scheme which is EM is expedient.

Corollary I

The rate of convergence of every EM scheme is controlled entirely by the parameter 'd'.

Proof : The rate of convergence of the above Markov process is determined by the eigenvalue of Q which is of magnitude less than unity[7]. Since one eigenvalue is always unity, the sum of diagonal trace of Q, indicates that the second eigenvalue is $1-(c_1+c_2)(1-d)$, which is only a function of 'd' and the penalty probabilities.

$= (1-c_1(1-d)) + (1-c_2(1-d)) - 1$
$= (c_1+c_2)(1-d) - 1$

Theorem III

The set of absolutely expedient schemes and the set of schemes which are EM are disjoint.

Proof : Lakshmivarahan and Thathachar[8] have proved the necessary and sufficient conditions for absolute expediency. These conditions do not permit the linear dependence of $\emptyset(.)$ and $\theta(.)$ which is a necessary condition for the scheme to be EM. Hence the theorem.

Remarks :

(1) The limiting value of $E[\underline{p}(\infty)]$ is exactly the same as that of the symmetric L_{RP} scheme.

(2) The L_{RP} scheme, which is the only EM scheme reported in the literature is obtained by substituting $d = w(.,.) = a$.

(3) One scheme which is of particular importance is the Generalized Linear EM (GL_{EM}) scheme. This is obtained when $w(.,.) = a$ and $d \geq a$. We shall study this scheme more extensively in the

next section and show how linear EM schemes can be designed to have an optimal limiting variance for a desired rate of convergence.

(4) Some examples of functions that can be used to design nonlinear EM schemes are :

a) $w(p_1,p_2) = a + bp_1p_2$

b) $w(p_1,p_2) = a + bp_1^K p_2^K$

In the first of these cases, it can be shown that the parameter 'b' can be used to change the variance of the limiting probabilities considerably for a fixed 'a'.

III VARIANCE OF EM SCHEMES

In this section we shall study some properties of the variance of schemes which are EM. In particular we shall study the class of EM schemes which are linear, and for which the variance can be minimized for a desired rate of convergence.

The Generalized Linear EM (GL_{EM}) scheme is obtained by using the function $w(.,.)=a$ in the updating scheme given in (1) and by using $d \geq a$. The resulting probability updating algorithm is given below for $i,j=1,2$ with $i \neq j$.

$$\begin{aligned} p_i(n+1) &= d - a\,p_j(n) && \text{if } a(n) = a_i \text{ and } b(n) = 1 \\ &= 1 - a\,p_j(n) && \text{if } a(n) = a_i \text{ and } b(n) = 0 \\ &= (1-d) + a\,p_i(n) && \text{if } a(n) = a_j \text{ and } b(n) = 1 \\ &= a\,p_i(n) && \text{if } a(n) = a_j \text{ and } b(n) = 0 \end{aligned}$$

Observe that the scheme defined by (1) reduces to the L_{EM} scheme if $d = a$. By studying the constraints on $\theta(.)$, it can be seen that this scheme will be of a Reward-Penalty nature if and only if $d=a$. However for all $d \geq a$, the scheme will be EM. Observe further that as proved in the previous section, the parameter 'a' has absolutely no control on the rate of convergence of the

scheme. The latter is controlled solely by 'd'. We now derive an explicit expression for the limiting variance of the GL_{EM} scheme.

Theorem IV

The GL_{EM} scheme has a limiting variance given by :

$$\text{Var}[p_i(\infty)] = \frac{c_1 c_2 [(1-a)^2 - 2(d-a)(1-d)(c_1+c_2)]}{(c_1 + c_2)^2 \quad [(1-a)^2 + 2a(c_1 + c_2)(1-d)]}$$

Proof : The distribution of $p_i(n+1)$ is as below.

$\phi(p_i) = d - \theta(p_j) = d - ap_j$

$p_i(n+1)$ =	$d - a\,p_j$	with prob. $p_i\,c_i$
=	$1 - a\,p_j$	with prob. $p_i\,(1-c_i)$
=	$(1-d) + a\,p_i$	with prob. $p_j\,c_j$
=	$a\,p_i$	with prob. $p_j\,(1-c_j)$

The conditional second moment of $p_1(n+1)$ is :

$$E[p_1^2(n+1) \mid p_1] = p_1^2 \{(c_1 + c_2)(d-1)2a + a(2-a)\}$$

$$+p_1 \{(c_1 + c_2)(d^2 - 2ad + 2a - 1) + c_2(1-d)2d + (1-a)^2\} + c_2(1-d)^2$$

Taking expectations on both sides and using the limiting value of $p_1(\infty)$ derived above, the result follows after considerable manipulation.

Remark : Note that the expression for the variance derived above reduces to the expression for the variance of the symmetric L_{RP} scheme when d=a.

Corollary II.1

For all $a,d \geq 0.5$, whenever the sum of the penalty probabilites is greater than 0.5 the variance of the limiting probabilities can be minimized for a given rate of convergence. The value of 'a' which minimizes the variance is given by a_{opt} below.

$$a_{opt} = \frac{1 - \sqrt{\dfrac{(1-d)^2\ (c_1+c_2)}{(1-d)^2(c_1+c_2) + 2(2d-1)}}}{1 + \sqrt{\dfrac{(1-d)^2\ (c_1+c_2)}{(1-d)^2(c_1+c_2) + 2(2d-1)}}}$$

Proof : The variance of $p_i(\infty)$ is minimized by differentiating the expression for the variance and setting the derivative to zero. In this case the first derivative is zero at a_{opt}. The second derivative can be guaranteed to be positive only when $a,d \geq 0.5$ and $(c_1 + c_2) \geq 0.5$.

Remarks :

(1) The superiority of the GL_{EM} scheme over the L_{EM} scheme is obvious. Whereas in the latter, the single parameter 'a' controlled both the rate of convergence and the limiting variance, in the former the rate of convergence can be chosen by appropriately choosing 'd' and the variance can be then appropriately controlled by varying 'a'. Note that whenever $(c_1 + c_2) \geq 0.5$ the parameters can be chosen to ensure that the limiting distribution has a minimum variance.

(2) An alternate approach to design the automaton would be to first decide on a value of 'a', and the value of 'd' which minimizes the variance is obtained by solving for d_{opt} as explained in the above corollary. The explicit form for d_{opt} in terms of 'a' is :

$$d_{opt} = \frac{\sqrt{(1-a)^2 + 2a(c_1+c_2)}}{(1-a) + \sqrt{(1-a)^2 + 2a(c_1+c_2)}}$$

(3) When $(c_1 + c_2) < 0.5$ there may or may not be a unique minimum variance. The expression for the derivative of the

variance is too complex and so nothing conclusive can be said. In such cases, the optimal design of the automaton would have to be by trial and error. Once the rate of convergence is chosen, the parameter 'a' can be varied between 0 and 'd' and the value of 'a' which minimizes the variance can be used.

(4) It may be noted that the optimum values for a and d depend on (c_1+c_2). This means additional knowledge of the environment but does not invalidate the leearning problem. For example, the value of (c_1+c_2) may be known but the better action may be unknown.

III.1 The Limiting Variance for the Nonlinear EM Scheme

In the general nonlinear EM scheme one has to know the exact form of the function w(.,.) to obtain an exact expression for the variance of the limiting probailities. As an example we consider the case when $w(p_1,p_2) = w_N = a + bp_1p_2$. The expression for the variance can be obtained by computing $E[p_1^2(n+1) \mid p_1]$ and using the limiting expected value of the probabilities as explained in the previous theorem.

To illustrate the power of the nonlinear scheme we compare the variance of this N_{EM} scheme with the variance of the L_{EM} scheme. In other words, we compare the limiting variances between the cases when $w_N=a+bp_1p_2$ and $w_L = a$. Let ΔVar be the difference between these variances.

$$\Delta Var = Var(N_{EM}) - Var(L_{EM})$$

Since both the schemes converge to the final expected value all the constant terms disappear. After much algebra it can be shown that

$$\Delta Var = Bp_1p_2$$

where, $B = \{ 2(c_1+c_2)(1-d)(w_N-w_L) + (w_N^2-w_L^2) - 2(w_N^2 \}$

If the parameter 'd' of the N_{EM} scheme is made equal to 'a', it can be shown that ΔVar is negative whenever

$$(c_1 + c_2) \leq 1 \quad \text{and} \quad 0 \leq b \leq 8(1-(c_1 + c_2)))(1-a)$$

or $$(c_1 + c_2) \geq 1 \quad \text{and} \quad 8(1-(c_1 + c_2)))(1-a) \leq b \leq 0.$$

In these situations the parameters of the automaton can be chosen to render the variance of the limiting probabilities less than the variance of the corresponding linear scheme.

In cases when $d \neq a$ no explicit expression can be obtained to satisfy $\Delta Var < 0$. In such cases one must resort to simulation to obtain the ideal values of 'a' and 'b' for a fixed rate of convergence.

IV EXPERIMENTAL RESULTS

IV.1 Simulation Results for the GL_{EM} Automaton

An environment with $c_1 = 0.2$ and $c_2 = 0.8$ was chosen for simulating the GL_{EM} automaton, and 'a' was chosen to be 0.6. The optimal value of 'd' for this value of a is 0.74.

To demonstrate the variation of $Var[p_i(\infty)]$ with 'd', we have plotted the value of $\hat{V}$ as a function of 'd' in Fig. I.

$$\hat{V} = 1/N \left[\sum_{j=1}^{N} (p_{1j}(\infty) - (c_2/(c_1+c_2))^2 \right]$$

In the above expression, N is the number of experiments, and $p_{1j}(\infty)$ is the final value of p_1 in the jth experiment. Note that we have used the exact value $c_2/(c_1 + c_2)$ in the computation instead of the sample mean of the final value. This is to avoid the errors that would be encountered by ignoring the effect of the variance of the sample mean. From the graph we observe that

the variance attains its minimum near the theoretically expected value of d=0.74.

To observe the effect of 'd' on the rate of convergence the quantity R_o has been plotted as a function of 'd' in Fig. II, where, R_o is the ratio of $E[p_1(1) - p_1(0)]$ to $E[p_1(\infty) - p_1(0)]$, where $p_1(0)=0.5$. Observe that R_o is the ratio of the rise of $E[p_1(n)]$ in the first step to the total rise that ought to be obtained. It can be shown that R_o is a function of 'd' and the penalty probabilities only. In this case, since the sum of the penalty probabilities is unity, the second eigenvalue is exactly equal to 'd'. Thus we expect R_o to decrease monotonically with 'd'. This monotonicity is clear from Fig. II.

III.2 Simulation Results for the Nonlinear EM Scheme

A nonlinear EM automaton was simulated to learn in three environments which had the sum of the penalty probabilities less than unity, equal to unity and greater than unity respectively. The actual penalty probabilities of the environments were:

(i)	$c_1 = 0.1$	$c_2 = 0.4$	$(c_1 + c_2) < 1$
(ii)	$c_1 = 0.2$	$c_2 = 0.8$	$(c_1 + c_2) = 1$
(iii)	$c_1 = 0.6$	$c_2 = 0.9$	$(c_1 + c_2) > 1$

The form of w(.,.) was as in the previous subsection $w_N=a+bp_1p_2$. In all these cases the parameter 'd' was made equal to 'a'.

The quantity R_o is defined as in the GL_{EM} scheme as the ratio of $(E[p_1(1)] - 0.5)$ to $(E[p_1(\infty)] - 0.5)$. A larger value of R_o implies that a larger rise in $E[p_1(n)]$ has occurred in the first step. In all these cases 400 experiments were conducted.

As in the GL_{EM} scheme, the sample variance of $p_1(\infty)$ was calculated by taking the average square deviations from the theoretical Expected limit.

The results of the experiments are highlighted below.

(1) In a given environment the value of R_o was almost the same irrespective of the values of 'b' ranging from -0.4 to 0.4. The values of this quantity (R_o) for the three environments were approximately 0.15, 0.4 and 0.1 respectively.

(2) The variation of the variance with the parameter 'b' has been plotted for the three environments in Fig.II. When $(c_1 + c_2) > 1$ the variance increased with 'b'. When $(c_1 + c_2) < 1$ the variance decreased with 'b'. Finally, in the case when $(c_1 + c_2) = 1$ the variance had a minimum when 'b' = 0, which corresponds to the L_{EM} scheme. These results were as expected.

IV. CONCLUSIONS

In this paper we have considered the general problem of designing stochastic learning automata in which the expected value of the action probabilities is the total state probability of an ergodic Markov chain. Automata which possess this property are said to be Ergodic in their Mean (EM). The only EM algorithm discussed in the literature is the symmetric Linear Reward-Penalty (L_{RP}) scheme.

We have considered the general problem of designing variable structure stochastic automata which are EM. The automata are fully defined by two probability updating functions $\emptyset(.)$ and $\theta(.)$. We have derived necessary and sufficient conditions on $\emptyset(.)$ and $\theta(.)$ that guarentee the scheme to be EM. The mean of the limiting distribution has been derived for the genral case. Some

results for the variance have been obtained for a typical nonlinear updating algorithm.

It has been shown that the set of absolutely expedient schemes is disjoint from the set of schemes that are EM.

In particular we have studied a whole family of linear schemes which are EM. Though these schemes are two parameter schemes, only one of these parameters controls the rate of convergence. The other parameter can be used to control the variance of the limitng distribution.

Simulation results have been included which highlight the theoretical results obtained for both linear and nonlinear schemes.

REFERENCES

1. Tsetlin, M.L., "On the Behaviour of Finite Automata in Random Media ", Automat. Telemekh., Vol.22, 1961, pp.1345-1354.

2. Tsetlin, M.L., "Automaton Theory and the Modelling of Biological Systems", New York and London, Academic, 1973.

3. Paz, A., "Introduction to Probabilistic Automata", New York, Academic, 1971.

4. Varshavskii, V.I., and Vorontsova, I.P., "On the behaviour of Stochastic Automata with Variable Structure", Automat. Telemek. (USSR), vol.24, 1963, pp.327-333.

5. Narendra, K.S., and Thathachar, M. A. L., Forthcoming Book on Learning Automata.

6. Narendra, K.S., and Thathachar, M. A. L., "Learning Automata -- A Survey",IEEE Trans. Syst. Man and Cybern., Vol. SMC-4, 1974, pp.323-334.

7. Isaacson, D.L., and Madson, R.W., "Markov Chains : Theory and Applications", Wiley, 1976.

8. Lakshmivarahan, S., and Thathachar, M.A.L., "Absolutely Expedient Algorithms for Stochastic Automata", IEEE Trans. on Syst. Man and Cybern., Vol.SMC-3, 1973, pp.281-286.

9. Norman, M.F., "Markov Processes and Learning Models", New York, Academic, 1972.

10. Norman, M.F., "Some Convergence Theorems for Stochastic Learning Models with Distance Diminshing Operators", J. Math. Psych., Vol. 5, 1968, pp.61-101.

11. Lakshmivarahan, S., "Learning Automata : Theory and Applications", Springer Verlag, New York, 1981.

12. Flerov, Y. A., "Some Classes of Multi Input Automata", Journal of Cybernetics", Vol. 2, 1972, pp. 112-122.

13. Tsypkin, Y. Z. and Poznyak, A. S., "Finite Learning Automata" Engineering Cybernetics, Vol. 10, 1972, pp. 478-490.

14. Poznyak, A. S., "Use of Learning Automata for the Control of Random Search", Automation and Remote Control, Vol. 33, No. 12, 1972, pp. 1992-2000.

15. El-Fatteh, Y. M., "Gradient Approach for Recursive Estimation and Control in Finite Markov Chains", Advances in Applied Prob., Vol. 13, 1981, pp. 778-803.

16. El-Fatteh, Y. M., "Multi Automaton Games : A Rationale for Expedient Collective Behavior",Systems and Control Letters, Vol. 1, 1982, pp. 332-339.

LIST OF FIGURES

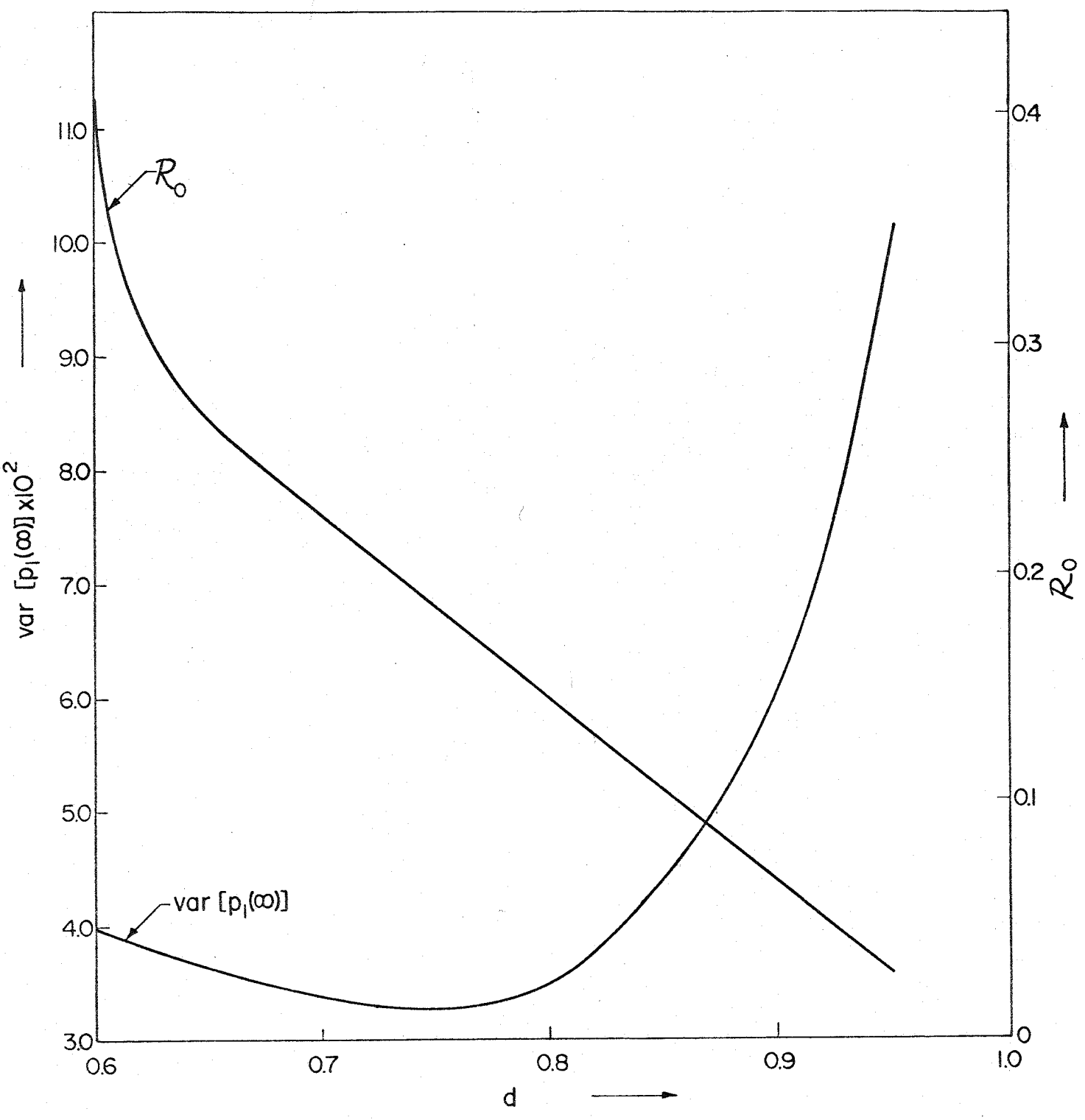

Figure 1: VARIATION OF SPEED AND VARIANCE WITH "d"

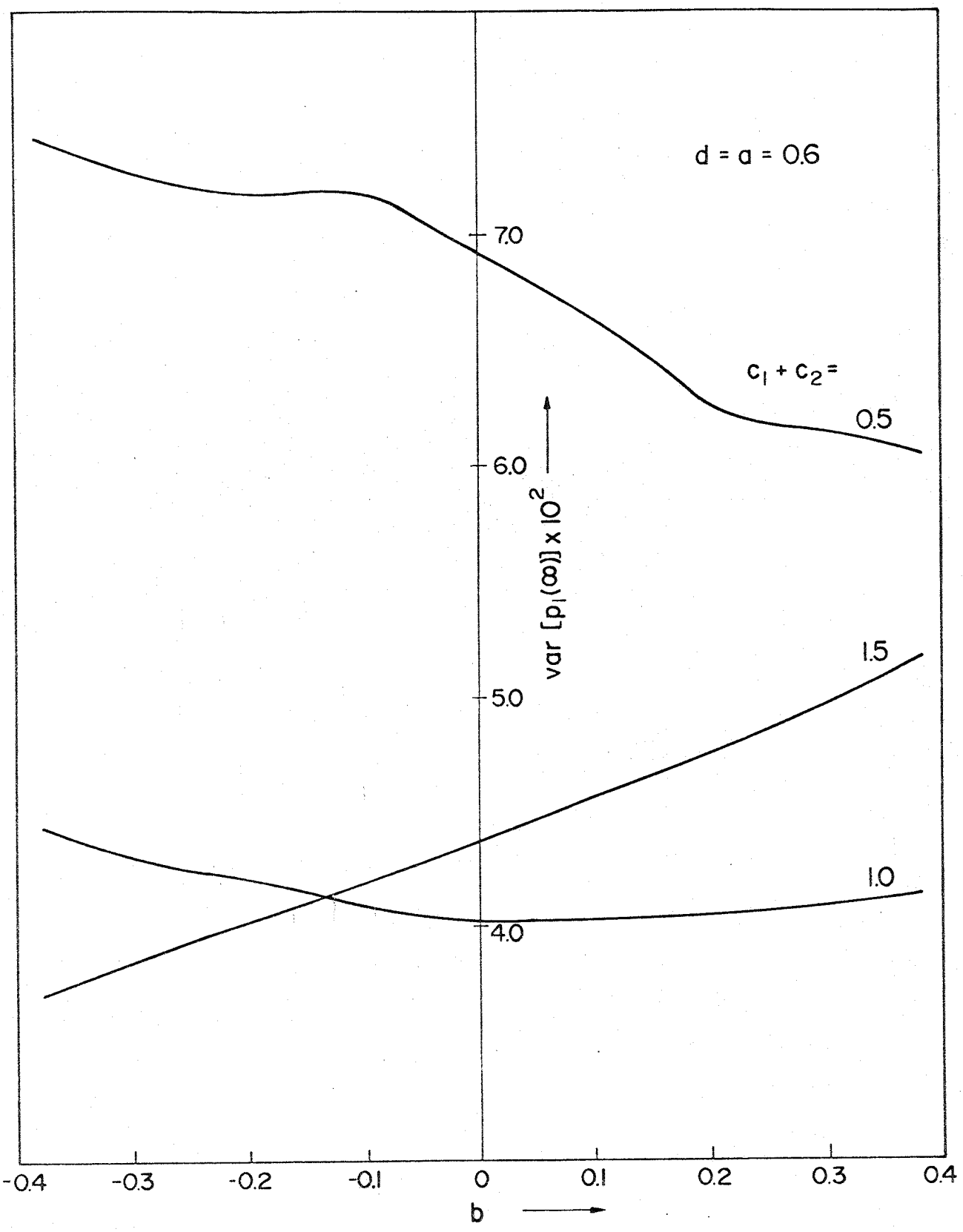

Figure II: N : VARIATION OF VARIANCE WITH "b"