MULTI-ACTION LEARNING AUTOMATA POSSESSING ERGODICITY OF THE MEAN

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

SCS TR-32

August 1983

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

** Department of Electrical Engineering, Indian Institute of Science, Bangalore 560012, India

A preliminary version of the paper was presented at the IASTED Conference on Measurement and Control, at Athens, September 2, 1983.

Partially supported by the Natural Sciences and Engineering Research Council of Canada

MULTI-ACTION LEARNING AUTOMATA POSSESSING ERGODICITY OF THE MEAN

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

ABSTRACT

Multi-action 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. Learning automata are said to possess Ergodicity of the Mean (EM) if the mean action probability is the state probablity (or unconditional probability) of an ergodic Markov chain. The only known algorithm which is Ergodic in the Mean (EM) is the Symmetric Linear Reward-Penalty (L_{RP}) scheme. Earlier [11] necessary and sufficient conditions have been derived for two-action nonlinear updating schemes to be Ergodic in the Mean (EM).

In this paper, we generalize the results of [11] and obtain necessary and sufficient conditions for the multi-action learning automaton to be Ergodic in the Mean (EM). The conditions involve two families of probability updating <u>functions</u>. It has been shown that for the automaton to be EM the two families must be linearly dependent. The vector defining the linear dependence is the only vector parameter which controls the rate of convergence of the automaton. Further, the technique for reducing the

---

+ School of Computer Science, Carleton University, Ottawa, Ontario, K1S 5B6, CANADA.
++ Department of Electrical Engineering, Indian Institute of Science, Bangalore, 560012, INDIA.

variance of the limiting distribution has also been discussed.

Just as in the two-action case, 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.

I INTRODUCTION

Automata models for learning have been used to model biological learning processes. The learning automaton is required to interact with an environment and to learn the optimal action which the environment offers. Such learning automata have had a variety of applications in parameter optimization, adaptive controlling of systems and in the routing of telephone calls.

The learning process of the automaton can be described as follows. Consider Fig. I. The environment with which the automaton interacts offers the latter a finite set of actions. The automaton is constrained to choose one of these actions. Once the action is chosen, the automaton is penalized by the environment, the penalty probability being dependent on the action chosen. A learning automaton is one which learns the action with the minimum penalty probability and which ultimately chooses this "more frequently" compared to the other actions.

Of the learning automata studied in the literature we are concerned with those which have transition matrices which are both time varying and stochastic. With no loss of generality, we assume that the output matrix is always deterministic [3]. Such automata are termed as Variable Structure Stochastic (VSS) automata. It can be shown that a VSS automata can be constructed by merely formulating a scheme by which the action probabilities can be updated.

An important class of VSS automata are those which possess ergodic properties. Ergodic VSS automata are known for their excellent learning properties when interacting with environments which have time varying penalty probabilities. Various ergodic schemes have been proposed and investigated by Lakshmivarahan [12], Flerov[13], Tsypkin and Poznyak[14,15] and El Fatteh [15,16].

The most simple 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 irrespective of the response of the environment. The limiting probability vector converges in distribution, and the form of this distribution has been known only for the symmetric version of the L_{RP} scheme which is a one-parameter probability updating algorithm.

To help understand the contributions of this paper we need the following definition introduced in [11].

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 was introduced by us in [11] and it appears important because it is one of the simple ways in which the mean of the action probability vector

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

can be made to possess certain desirable characterstics. All the well studied properties of an ergodic Markov chain such as the limiting distribution, the rate of convergence etc. can be readily applied to the learning automaton possessing this property. The literature reports only one scheme that is known to be EM and this is the symmetric L_{RP} scheme.

In [11] we considered the general problem of the two-action probability updating scheme possessing ergodicity of the mean. The updating algorithm was given in terms of two nonlinear functions $\emptyset(.)$ and $\Theta(.)$. Two necessary and sufficient conditions involving these functions were derived for the scheme to be EM. The first of these conditions resembles the one proven to be necessary and sufficient for absolute expediency [5,6,8], and the second is a linear constraint involving the functions and a constant. The latter constant is the <u>only</u> parameter which controls the rate of convergence of the scheme. Further, it was shown in [11] that 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 was also proposed.

In this paper we consider the problem of the multi-action learning automaton being EM. For the R-action environment, the updating scheme is defined using two families of functions $\{\emptyset_i(.) \mid i=1,...,R\}$ and $\{\Theta_i(.) \mid i=1,...,R\}$. These functions are explicit functions of the action probability vector. We refer to these families of functions as $\{\emptyset(.)\}$ and $\{\Theta(.)\}$ respectively.

The highlights of the contributions of this paper are that the necessary and sufficient conditions on $\{\phi(.)\}$ and $\{\theta(.)\}$ have been derived which render the automaton EM. These conditions can be viewed as vector versions of the corresponding conditions imposed in the two-action problem. Further, we show that the scheme can be EM if and only if $\phi_i(.)$ and $\theta_i(.)$ are linearly dependent. The vector of coefficients which specify the linear dependence has been shown to be the ONLY set of parameters which influence the rate of convergence of the learning automaton.

We have also suggested a technique by which the variance of the limiting action probabilities can be minimized.

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 and prove some fundamental theorems regarding the rate of convergence of EM schemes and of the limiting action probabilities. 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:

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

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 defined below is referred to as the penalty probability.

$c_i = \Pr [b(n) = 1 \mid a(n) = a_i] \qquad (i = 1,\ldots,R)$

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.

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

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</u> <u>expedient</u> if

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

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

I.2 The R-action L_{EM} Scheme

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

$$
\begin{aligned}
p_i(n+1) &= a p_i && \text{if } a(n) = a_j \text{ and } b(n) = 0 \\
&= 1 - a \sum_{j \neq i} p_j && \text{if } a(n) = a_i \text{ and } b(n) = 0 \\
&= b p_i && \text{if } a(n) = a_i \text{ and } b(n) = 1 \\
&= b p_i + \frac{1-b}{R-1} && \text{if } a(n) = a_j \text{ and } b(n) = 1
\end{aligned}
$$

To simplify the notation, unless explicitly stated we use p_i to refer to the probability $p_i(n)$. The vector $\underline{p}$ will refer to $[p_1, p_2, \ldots, p_R]^T$. Note that if the action a_i is chosen and a penalty is obtained, the decrease in probability is shared among the rest. In this form of the L_{RP} scheme $E\,[p_i\,(n+1) \mid \underline{p}]$ has the expression:

$$
E\,[p_i\,(n+1) \mid \underline{p}] = (b-a)\, p_i \sum_{j=1}^{R} p_j\, c_j + p_i\,(1 - c_i + a\, c_i) + \frac{1-b}{R-1} \sum_{j \neq i} p_j\, c_j
$$

Observe that $E\,[p_i(n+1)]$ is not linear in $\underline{p}$. It consists of a sum of terms quadratic in $p_i p_j$. Because of this, the form of limiting distribution of the general L_{RP} scheme is unknown. However, in the symmetric case when b = a, the quadratic terms disappear, yielding the vector equality

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

where the stochastic matrix A has elements

$$A_{ii} = 1 - (1 - a)\, c_i$$

$$A_{ji} = (1 - a)\, \frac{c_j}{R-1}$$

It can be shown that since E [$\underline{p}(n)$] possesses the above Markov property, the limiting value of the expected action probabilities are

$$E\,[p_i(\infty)] = \frac{1/c_i}{\sum_{j=1}^{R} 1/c_j}$$

The limiting expected penalty is thus the harmonic mean of the individual penalty probabilities. Since the harmonic mean is always less than the arithmetic mean, the R-action symmetric L_{RP} is expedient in all environments. Since the R-action symmetric L_{RP} scheme is Ergodic in the Mean, we shall refer to it as the L_{EM} scheme. Currently, this is the only R-action scheme known to be EM.

We now study generalized nonlinear EM automaton.

II NONLINEAR SCHEMES ERGODIC IN THE MEAN

We shall first consider the general problem of designing nonlinear EM learning schemes. Two sets of necessary and sufficient conditions for probability updating schemes to be EM have been derived. The conditions involve two families of arbitrary functions $\emptyset_i(.)$ and $\theta_i(.)$ defined for i=1,...,R. The first set of conditions is similar to the conditions required to guarentee absolute expediency [5,6,8,12]. The second set constrains the functions $\theta_i(.)$ and $\phi_i(.)$ to be linearly dependent.

The probability updating scheme for R-actions is given below.

$$
\begin{aligned}
p_j(n+1) &= \phi_j(\underline{p}) && \text{if } a(n) = a_i \quad b(n) = 1 \\
&= 1 - \sum_{i \neq j} \phi_i(\underline{p}) && \text{if } a(n) = a_j \quad b(n) = 1 \\
&= \theta_j(\underline{p}) && \text{if } a(n) = a_i \quad b(n) = 0 \\
&= 1 - \sum_{i \neq j} \theta_i(\underline{p}) && \text{if } a(n) = a_j \quad b(n) = 0
\end{aligned}
\tag{1}
$$

The updating scheme is easily comprehended. If the action chosen is a_i and a penalty is obtained, the probability p_j is updated to $\phi_j(\underline{p})$. Once all the other action probabilities have been updated, the action probability of the action chosen is set to render the sum of the probabilities to be unity.

In a similar way, if a(n) is a_i and the response is a reward, the algorithm updates all the other action probablities to $\theta_i(\underline{p})$ for all $i \neq j$. Again, $p_i(n+1)$ is calculated so that the sum of the action probabilities is unity.

Note that for the scheme to be strictly of a Reward-Penalty nature the following obvious inequalities must hold for all $j=1,\ldots,R$.

$$p_j \leq \phi_j(\underline{p}) \leq 1$$

$$0 \leq \theta_j(\underline{p}) \leq p_j$$

We now present some properties of the generalized nonlinear EM scheme.

<u>Theorem I:</u>

The sufficient and necessary conditions of the probability updating scheme defined by (1) to be EM are:

$$\frac{\theta_1(\underline{p})}{p_1} = \frac{\theta_2(\underline{p})}{p_2} = \cdots = \frac{\theta_R(\underline{p})}{p_R} = L$$

$$\text{and} \qquad \theta_i(\underline{p}) - \phi_i(\underline{p}) = d_i$$

Proof: By virtue of the updating scheme defined by (1), $p_j(n+1)$ has the following distribution:

$$\begin{aligned} p_j(n+1) &= \phi_j(\underline{p}) && \text{w.p.} \quad \sum_{i \neq j} p_i c_i \\ &= 1 - \sum_{i \neq j} \phi_i(\underline{p}) && \text{w.p.} \quad p_j c_j \\ &= \theta_j(\underline{p}) && \text{w.p.} \quad \sum_{i \neq j} p_i (1-c_i) \\ &= 1 - \sum_{i \neq j} \theta_i(\underline{p}) && \text{w.p.} \quad p_j (1-c_j) \end{aligned}$$

To simplify the notation, we shall omit the arguments for $\phi_i(\underline{p})$ and $\theta_i(\underline{p})$ observing they are always $\underline{p}$.

$$\begin{aligned} E[p_j(n+1)/\underline{p}] &= c_j p_j \{\sum_{i \neq j} (\theta_i - \phi_i)\} + \{\sum_{i \neq j} (\phi_j - \theta_i)\} p_i c_i \\ &\quad + p_j \{1 - \sum_{i \neq j} \theta_i\} + \theta_j (1-p_j) \\ &= c_j p_j \{\sum_{i \neq j} (\theta_i - \phi_i)\} + \{\sum_{i \neq j} (\phi_j - \theta_j) p_i c_i\} \\ &\quad + p_j \{1 - \sum_{i=1}^{R} \theta_i\} + \theta_j \end{aligned}$$

Since the first two terms of the above involve the penalty probabilities, if $E[p_i(n+1)]$ is to be a linear function of $E[\underline{p}(n)]$ each quantity in the parenthesis of these terms must be a constant. This is a consequence of the fact that cancellations between the first and second terms cannot occur because the updating functions cannot be explicit functions of the <u>unknown</u> penalty probabilities $\{c_i\}$. Hence, <u>a</u> set of necessary and sufficient conditions for the scheme to be EM is

$$\theta_i - \phi_i = d_i \qquad \text{for } i=1,\ldots,R.$$

Consider the last two terms

$$p_j \{ 1 - \sum_{i=1}^{R} \theta_i \} + \theta_j$$

We contend that these terms are linear in $\underline{p}$ if and only if

$$\frac{\theta_1}{p_1} = \frac{\theta_2}{p_2} = \quad = \quad = \frac{\theta_R}{p_R} \qquad (3)$$

Clearly, if (3) is enforced, $\theta_j = p_j \sum_{i=1}^{R} \theta_i$, and hence

$$p_j - p_j \sum_{i=1}^{R} \theta_i + \theta_j = p_j. \qquad (4)$$

Hence (3) is obviously a sufficient constraint.

We prove necessity of (3) by observing that the most general form of a linear function in $\underline{p}$ is $\sum_{i=1}^{R} a_i p_i$. For the last two terms to be linear in $\underline{p}$,

$$p_j - p_j \sum_{i=1}^{R} \theta_i + \theta_j = a_j p_j + L(\underline{p}) \qquad (5)$$

where L is linear in p_i, $i \neq j$, $i=1,\ldots,R$.

Summing (5) over all values of j, yields the LHS to be

$$\sum_{j=1}^{R} p_j - (\sum_{j=1}^{R} p_j)(\sum_{i=1}^{R} \theta_i) + (\sum_{j=1}^{R} \theta_j)$$

which is unity.

The RHS of (5) is unity if and only if L is identically zero and if a_j is identically unity. Hence (3) is necessary for the system to be EM. Hence the theorem.

Remark : The linear constraint involving $\emptyset(.)$ and $\theta(.)$ is

$$\emptyset_i(\underline{p}) - \theta_i(\underline{p}) = d_i$$

Since the penalty probabilities are unknown, with no apriori information there is no loss in generality by assuming that the constants d_i are all equal for $i=1,\ldots,R$. If we use

$$-d_i = \frac{1 - d}{R - 1}$$

we obtain the relationship obeyed by the expected value of the action probabilities as

$$E\,[p_j(n+1)] = E\,[p_j(n)]\ \{1 - (1-d)\ c_j\} + \{\frac{1-d}{R-1} \sum_{i \neq j} c_i\}.\ E[p_i(n)]$$

In matrix form, $E\,[\underline{p}(n+1)] = A^T\ E\,[\underline{p}(n)]$ where,

$$A = \begin{bmatrix} 1-(1-d)c_1 & \frac{(1-d)c_1}{R-1} & \cdots & & \cdots & \frac{(1-d)c_1}{R-1} \\ \frac{(1-d)c_2}{R-1} & 1-(1-d)c_2 & & & & \frac{(1-d)c_2}{R-1} \\ \frac{(1-d)c_3}{R-1} & & 1-(1-d)c_3 & \cdots & & \\ & & & & & \\ \frac{(1-d)c_R}{R-1} & & & & \frac{(1-d)c_R}{R-1} & 1-(1-d)c_R \end{bmatrix}$$

We now prove a theorem concerning the rate of convergence of the limiting vector.

Theorem II

The rate of convergence of a nonlinear EM scheme is determined entirely by the set of parameters $\{d_i \mid i=1,\ldots R\}$ which relate $\emptyset_i(.)$ and $\theta_i(.)$.

Proof: Subject to the conditions specified by Theorem I the expected value of the action probabilities obeys the martix equation specified above. The matrix A is Markovian. Hence, the rate of convergence of this Markov chain is controlled by the eigenvalue of A (other than unity) of largest magnitude. The latter is a function only of the d_i's and not of the functions $\emptyset_i(.)$ and $\theta_i(.)$. Hence the theorem.

For the rest of this paper we shall assume that the d_i's are all equal. In any particular problems, if there is reason to prefer one action over the other, the d_i's will be distinct. In such a case the matrix relationship $E[\underline{p}(n+1)] = A^T E[\underline{p}(n)]$ will still be obeyed except that the matrix A will involve the set of parameters, d_i, as opposed to a single parameter d.

Theorem III

In the case when the d_i's are all equal, the limiting expected action probabilities are all independent of d and have the value,

$$p_i^* = \frac{\frac{1}{c_i}}{\sum_{i=1}^{R} \frac{1}{c_i}} \qquad i=1,\ldots,R$$

Proof To get the limiting expected value of $\underline{p}(.)$ we solve

$$\underline{p}^* = A^T \underline{p}^*$$

p^* is thus the eigenvector of the eigenvalue which is unity.

Solving, $[I-A]^T\underline{p}^* = 0$ yields for the first row,

$$(1-d)\{c_1 p_1^* - \frac{1}{R-1} \sum_{i=2}^{R} c_i p_i^*\} = 0$$

whence $p_1^* = \frac{J}{Rc_1}$

where J is a constant independent of 'i' and is equal to

$\sum_{i=1}^{R} p_i^* c_i$. Similarly,

$$p_i^* = \frac{J}{Rc_i}$$

Since, $\sum_{i=1}^{R} p_i^*$ is unity, $J = \frac{R}{\sum_{i=1}^{R} \frac{1}{c_i}}$

Hence, $p_i^* = \frac{\frac{1}{c_i}}{\sum_{i=1}^{R} \frac{1}{c_i}}$

Corollary I: The Generalized Nonlinear EM scheme with d_i being equal for all the actions is expedient.

Proof: The result is proved from the above theorem by observing that the harmonic mean of a sequence of numbers is never less than that arithmetic mean.

Remarks:

(1) The symmetric L_{RP} scheme is obtained by using $\theta_i(\underline{p})=ap_i$ and d=a for i=1,...,R. Observe that the limiting value of the expected action probabilities from Theorem III above is identical to the limited value of the correspondign case in the symmetric

L_{RP} scheme.

(2) An example of a nonlinear function which can be used for $\theta_j(.)$ is

$$\frac{\theta_j}{p_j} = a + bp_1p_2...p_R = L$$

When b=0 and d≠a the scheme obtained is

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

Note that this is a two parameter updating scheme which is EM, as opposed to the only scheme possible in the format of the L_{RP} scheme described in Section III. For this scheme to be of a reward-penalty nature the parameter d must equal a. The distinctiveness of the scheme lies in the fact that the scheme is EM and yet has two parameters one of which solely controls the rate of convergence and the second parameter, 'a', can be used to independently minimize the variance.

We conclude this section by observing that the set of automata which are EM is disjoint from the set of automata which are absolutely expedient.

Theorem IV:

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

Proof: Lakshmivarahan and Thathachar [6,8,12] have proved the necessary and sufficient conditions for absolute expediencey. These conditions do not permit the linear dependence of $\emptyset(.)$ and $\theta(.)$ which is a necessary and sufficient condition for the scheme to be EM. Hence the theorem.

III DESIGN CONSIDERATIONS

Nonlinear EM automata can be designed using functions of the form specified by Remark (ii) above. if the penalty probabilities are known (though the actions to which they belong are unknown) the design process is rendered more easy. A suitable value of 'd' can be chosen so that the eigenvalues of the resulting transition matrix are determined by the convergence requirements.

For the two action case expressions have been derived for the values of the parameters which minimize the variance of the limiting action probabilities. In the R-action case no such expressions are available. The rest of the parameters in the scheme are determined by trial and error with the intention of minimizing the limiting variance. To demonstrate how this is done we study an environment with penalty probabilities:

$c_1 = 0.65$ $\quad c_2 = 0.2$ $\quad c_3 = 0.5$

$c_4 = 0.4$ $\quad c_5 = 0.85$

Observe that a_2 is the optimal action and this action is chosen asymptotically with an expected probability p_2^*, where,

$$p_2^* = \frac{(1/0.2)}{(1/0.65) + (1/0.2) + (1/0.5) + (1/0.4) + (0.85)} = 0.40394$$

The Nonlinear scheme which was used had the following form:

$$\frac{\theta_j}{p_j} = a + bp_1p_2p_3p_4p_5$$

and $\phi_j(\underline{p}) - \theta_j(\underline{p}) = d$

To simplify matters, d was set equal to a = 0.6. To study the variation of the limiting variance with 'b' we have plotted the value of V^+ as a function of 'b', where,

$$V^+ = \frac{1}{N} \sum_{j=1}^{N} (p_{2j}(\infty) - p_2^*)^2$$

In the above expression, N is the number of experiments, and $p_{2j}(\infty)$ is the final value of p_2 in the jth experiment. Note that we have used the exact value of p_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 Fig.II we observe the variation of V^+ with respect to 'b'. The value of the variance seems to be minimized when b is nearly 500. Observe that this variance is less than the variance of the corresponding L_{EM} scheme obtained when b=0.

If we keep both 'a' and 'b' as parameters to be varied their optimal values which reduce the variance even further can be obtained. As opposed to the two-action EM schemes [11] due to the complexity of the expressions inolved we have been unable to obtain an explicit relationship for the limiting variance. We have thus to resort to simulation to get the most desirable parameters. The problem of deriving a closed form expression for the variance for the family of linear and nonlinear EM schemes remains an unsolved problem.

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 multi-action variable structure stochastic automata which are EM. The automata are fully defined by two families of probability updating functions $\phi_i(.)$ and $\theta_i(.)$. We have derived necessary and sufficient conditions on $\phi_i(.)$ and $\theta_i(.)$ that guarentee the scheme to be EM. The conditions on $\phi_i(.)$ and $\theta_i(.)$ require that they be linearly dependent. Further, the nonlinear part of these functions must obey a simple relationship which is similar to the conditions derived for the two-action EM automata [11].

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 limiting distribution.

Simulation results have been included which highlight the strategy to be followed in the process of designing nonlinear EM automata.

ACKNOWLEDGEMENTS: The authors would like to express their sincere appreciation to Professor D. Dawson of the Department of Mathematics, Carleton University, Ottawa, for some invaluable discussions. It was one of these discussions which helped to solve a crucial problem encountered while proving Theorem I.

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. Thathachar, M.A.L., and Oommen, B.J., " Learning Automata Possessing Ergodicity of the Mean: The Two Action Case", To appear in the IEEE Trans. for Syst. Man and Cybern. (Issue not known).

12. Lakishmivarahan, S., "Learning Algorithms Theory and Applications", Springer Verlag, New York, 1981.

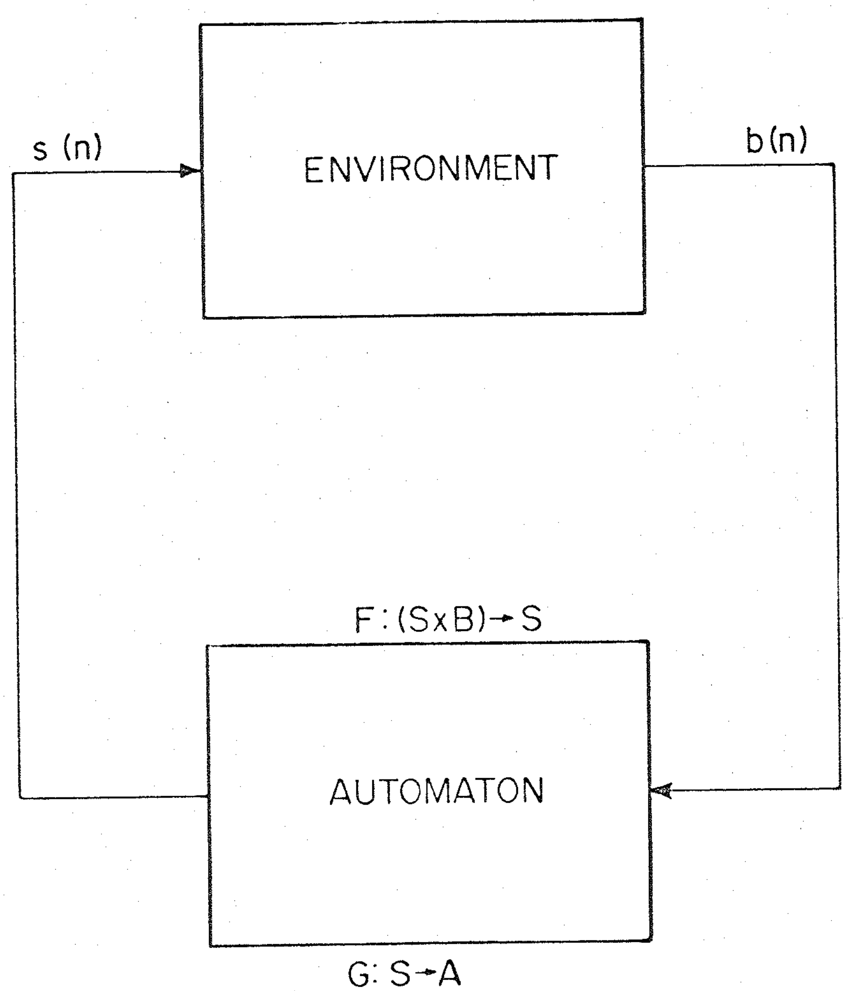
13. Flerov, Y. A., "Some Classes of Multi Input Automata", Journal of Cybernetics", Vol. 2, 1972, pp. 112-122.

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

15. 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.

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

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



$b(n) \in \{0, 1\} = B$

$s(n) \in \{s_1, s_2, \ldots s_N\} = S$

$a(n) \in \{a_1, a_2, \ldots a_R\} = A$

FIG. 1: THE AUTOMATON-ENVIRONMENT INTERACTION

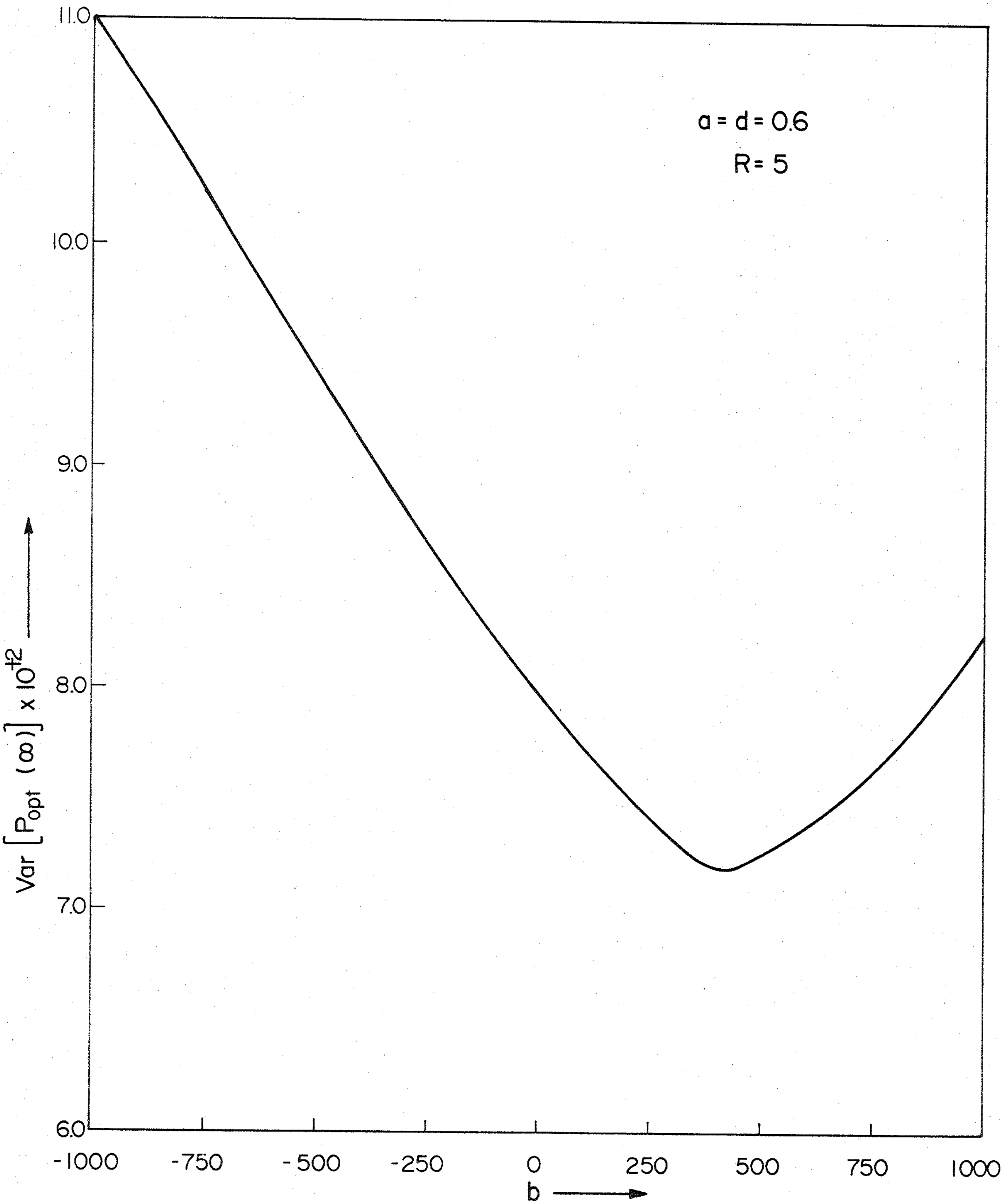


FIGURE II: NONLINEAR EM SCHEME VARIATION OF VARIANCE WITH "b"

CARLETON UNIVERSITY
School of Computer Science

Bibliography of SCS Reports

SCS-TR-1 THE DESIGN OF CP-6 PASCAL
Jim des Rivieres and Wilf R. LaLonde, June 1982.

SCS-TR-2 SINGLE PRODUCTION ELIMINATION IN LR(1) PARSERS:A SYNTHESIS
Wilf R. LaLonde, June 1982.

SCS-TR-3 A FLEXIBLE COMPILER STRUCTURE THAT ALLOWS DYNAMIC PHASE ORDERING
Wilf R. LaLonde and Jim des Rivieres, June 1982.

SCS-TR-4 A PRACTICAL LONGEST COMMON SUBSEQUENCE ALGORITHM FOR TEXT COLLATION
Jim des Rivieres, June 1982.

SCS-TR-5 A SCHOOL BUS ROUTING AND SCHEDULING PROBLEM
Wolfgang Lindenberg, Frantisek Fiala, July 1982.

SCS-TR-6 ROUTING WITHOUT ROUTING TABLES
Nicola Santoro, Ramez Khatib, July 1982.

SCS-TR-7 CONCURRENCY CONTROL IN LARGE COMPUTER NETWORKS
Nicola Santoro, Hasan Hural, July 1982.

SCS-TR-8 ORDER STATISTICS ON DISTRIBUTED SETS
Nicola Santoro, Jeffrey B. Sidney, July 1982.

SCS-TR-9 OLIGARCHICAL CONTROL OF DISTRIBUTED PROCESSING SYSTEMS
Moshe Krieger, Nicola Santoro, August 1982.

SCS-TR-10 COMMUNICATION BOUNDS FOR SELECTION IN DISTRIBUTED SETS
Nicola Santoro, Jeffrey B. Sidney, September 1982.

SCS-TR-11 A SIMPLE TECHNIQUE FOR CONVERTING FROM A PASCAL SHOP TO C SHOP
Wilf R. LaLonde, John R. Pugh, November 1982.

SCS-TR-12 EFFICIENT ABSTRACT IMPLEMENTATIONS FOR RELATIONAL DATA STRUCTURES
Nicola Santoro, December 1982.

SCS-TR-13 ON THE MESSAGE COMPLEXITY OF DISTRIBUTED PROBLEMS
Nicola Santoro, December 1982.

SCS-TR-14 A COMMON BASIS FOR SIMILARITY MEASURES INVOLVING TWO STRINGS
R.L. Kashyap and B.J. Oommen, January 1983.

SCS-TR-15 SIMILARITY MEASURES FOR SETS OF STRINGS
R.L. Kashyap and B.J. Oommen, January 1983.

SCS-TR-16 THE NOISY SUBSTRING MATCHING PROBLEM
R.L. Kashyap and B.J. Oommen, January 1983.

SCS-TR-17 DISTRIBUTED ELECTION IN A CIRCLE WITHOUT A GLOBAL SENSE OF ORIENTATION
E. Korach, D. Rotem, N. Santoro, January 1983.

.../2

SCS-TR-18 A GEOMETRICAL APPROACH TO POLYGONAL DISSIMILARITY AND THE CLASSIFICATION OF CLOSED BOUNDARIES
R.L. Kashyap and B.J. Oommen, January 1983.

SCS-TR-19 SCALE PRESERVING SMOOTHING OF POLYGONS
R.L. Kashyap and B.J. Oommen, January 1983.

SCS-TR-20 NOT-QUITE-LINEAR RANDOM ACCESS MEMORIES
Jim des Rivieres, Wilf LaLonde and Mike Dixon, August 1982, Revised March 1, 1983.

SCS-TR-21 SHOUT-ECHO SELECTION IN DISTRIBUTED FILES
D. Rotem, N. Santoro, J. B. Sidney, March 1983.

SCS-TR-22 DISTRIBUTED RANKING
E. Korach, D. Rotem, N. Santoro, March 1983

SCS-TR-23 A REDUCTION TECHNIQUE FOR SELECTION IN DISTRIBUTED FILES : I
N. Santoro, J. B. Sidney, April 1983.

SCS-TR-24 LEARNING AUTOMATA POSSESSING ERGODICITY OF THE MEAN : THE TWO ACTION CASE
M.A.L. Thathachar and B. J. Oommen, May 1983

SCS-TR-25 ACTORS - THE STAGE IS SET .
John R. Pugh, June 1983

SCS-TR-26 ON THE ESSENTIAL EQUIVALENCE OF TWO FAMILIES OF LEARNING AUTOMATA
M. A. L. Thathachar and B. J. Oommen, May 1983

SCS-TR-27 GENERALIZED KRYLOV AUTOMATA AND THEIR APPLICABILITY TO LEARNING IN NONSTATIONARY ENVIRONMENTS
B. J. Oommen, June 1983

SCS-TR-28 ACTOR SYSTEMS: SELECTED FEATURES
Wilf R. LaLonde, July 1983.

SCS-TR-29 ANOTHER ADDENDUM TO KRONECKER'S THEORY OF PENCILS
M. D. Atkinson, August 1983

SCS-TR-30 SOME TECHNIQUES FOR GROUP CHARACTER REDUCTION
M. D. Atkinson and R. A. Hassan, August 1983

SCS-TR-31 AN OPTIMAL ALGORITHM FOR GEOMETRICAL CONGRUENCE
M. D. Atkinson, August 1983

SCS-TR-32 MULTI-ACTION LEARNING AUTOMATA POSSESSING ERGODICITY OF THE MEAN
B. J. Oommen and M. A. L. Thathachar, August 1983