INFORMATION SCIENCES **35**, 183–198 (1985)

Multiaction Learning Automata Possessing Ergodicity of the Mean

B. J. OOMMEN

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

and

M. A. L. THATHACHAR

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

ABSTRACT

Multiaction 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 according to whether the environment responds with a reward or a penalty. Learning automata are said to possess *ergodicity of the mean* if the mean action probability is the state probability (or unconditional probability) of an ergodic Markov chain. In an earlier paper [11] we considered the problem of a two-action learning automaton being ergodic in the mean (EM). The family of such automata was characterized completely by proving the necessary and sufficient conditions for automata to be EM. In this paper, we generalize the results of [11] and obtain necessary and sufficient conditions for the multiaction learning automaton to be EM. These conditions involve two families of probability updating *functions*. It is 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 variance of the limiting distribution is discussed. Just as in the two-action case, it is 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 the routing of telephone calls.

The learning process of the automaton can be described as follows. Consider Fig. 1. The environment with which the automaton interacts offers the latter a

52 Vanderbilt Ave., New York, NY 10017 0020-0255/85/$03.30

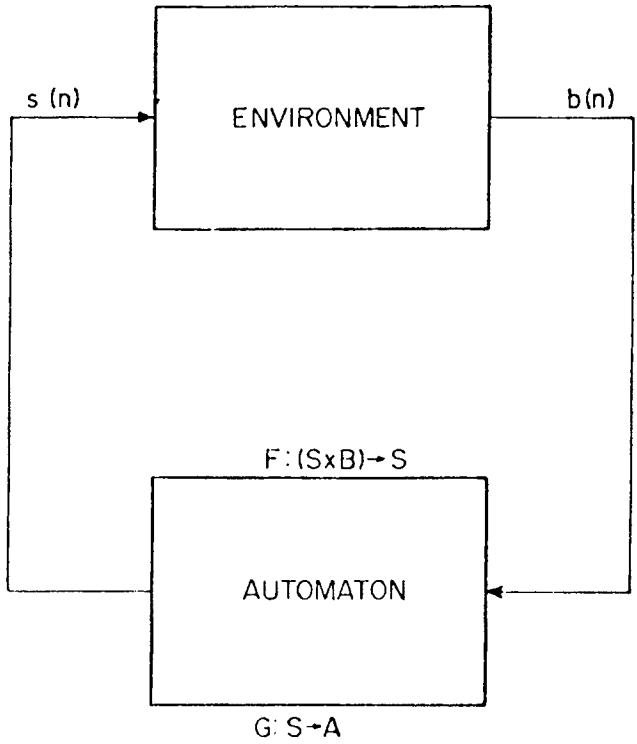

Fig. 1. The automaton-environment interaction: $b(n) \in \{0,1\} = B$, $s(n) \in \{s_1, s_2, \ldots, s_N\} = S$, $a(n) \in \{a_1, a_2, \ldots, a_R\} = A$.

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 (in some sense) than 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 variable-structure stochastic (VSS) automata. It can be shown that a VSS automaton 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 proper-

ties 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 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 irrespective of the response of the environment. The limiting probability vector converges in distribution, and the form of this distribution is at present 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[1] of an ergodic Markov chain.

REMARK. Consider the mean action probability of the automaton at the nth time instant. This vector is either stochastically independent of the corresponding vector at the previous time instants, or dependent on the values it took earlier. The former assumption (namely, that of independence) is obviously meaningless in the framework of a learning system. To investigate the question of the dependence of this vector on its history, the concept of ergodicity of the mean (EM) was introduced in [11]. Observe that such a study is productive, inasmuch as the fact remains that an ergodic Markov chain is probably one of the simplest ways of describing the dependence of two random vectors. Indeed, ergodicity of the mean is a powerful characterization of the set of ergodic automata. Further, such a characterization permits the use of many well-known techniques involved in the study of ergodic Markov chains, namely those by which the limiting distribution and the rate of convergence can be studied.

Although ergodicity of the mean is an interesting (and possibly one of the simplest) ways by which a learning automaton can be characterized, the only known algorithm possessing this property is the one known in the literature, as the symmetric linear reward-penalty (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 $\phi(\cdot)$ and $\theta(\cdot)$. Two necessary and sufficient conditions involving these functions were derived for the scheme to be EM. The first of

[1]Also called "absolute" or "unconditional" probability.

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 *only* 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 corresponding L_{RP} automaton was also proposed.

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

The main contribution of this paper are as follows: Necessary and sufficient conditions on $\{\phi(\cdot)\}$ and $\{\theta(\cdot)\}$ 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 $\phi_i(\cdot)$ and $\theta_i(\cdot)$ 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. Finally we present simulation results which demonstrate the learning capabilities of the automata discussed.

1.1. FUNDAMENTALS

The automaton selects an action $a(n)$ at a time instant n. Here $a(n)$ is any one of a finite set $(a_1,\ldots,a_R)$ and is selected on the basis of an $R\times 1$ probability vector $\mathbf{p}(n)$ whose components are

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

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 defined below is referred to as the penalty probability:

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

Thus the environment is characterized by the set of penalty probabilities. The automaton updates the vector $\mathbf{p}(n)$ 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 *a priori* information, the automaton chooses the actions with equal probability. The expected penalty is thus initially

$$M_0 = \sum_{i=1}^{R} p_i(0)\, c_i = \frac{1}{R} \sum_{i=1}^{R} c_i \qquad [\text{since} \quad p_i(0) = 1/R].$$

An automaton is said to learn *expediently* if, as time tends towards infinity, the expected penalty is less than M_0. The automaton is *absolutely expedient* if

$$E[\, M(n+1) \,|\mathbf{p}(n)] < M(n)$$

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:

$$p_i(n+1) = \begin{cases} ap_i & \text{if} \quad a(n) = a_j \text{ and } b(n) = 0, \\ 1 - a \sum_{j \neq i} p_j & \text{if} \quad a(n) = a_i \text{ and } b(n) = 0, \\ bp_i & \text{if} \quad a(n) = a_i \text{ and } b(n) = 1, \\ bp_i + \dfrac{1-b}{R-1} & \text{if} \quad a(n) = a_j \text{ and } b(n) = 1. \end{cases}$$

To simplify the notation, unless explicitly stated we use p_i to refer to the probability $p_i(n)$. The vector $\mathbf{p}$ 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 $E[p_i(n+1)|\mathbf{p}]$ has the expression

$$E[p_i(n+1)|\mathbf{p}] = (b-a)\, p_i \sum_{j=1}^{R} p_j c_j + p_i(1 - c_i + ac_i)$$

$$+ \frac{1-b}{R-1} \sum_{j \neq i} p_j c_j .$$

Observe that $E[p_i(n+1)]$ is not linear in $\mathbf{p}$. 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

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

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 $E[\mathbf{p}(n)]$ possesses the above Markov property, the limiting value of the expected action probabilities are

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

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 automata.

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 $\phi_i(\cdot)$ and $\theta_i(\cdot)$ defined for $i = 1, \ldots, R$. The first set of conditions is similar to the conditions required to guarantee absolute expediency [5, 6, 8, 12]. The second set constrains the functions $\theta_i(\cdot)$ and $\phi_i(\cdot)$ to be linearly dependent.

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

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

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 $\phi_j(\mathbf{p})$. 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 probabilities to $\theta_i(\mathbf{p})$ 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:

$$p_j \leqslant \phi_j(\mathbf{p}) \leqslant 1,$$

$$0 \leqslant \theta_j(\mathbf{p}) \leqslant p_j$$

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

THEOREM I. *Sufficient and necessary conditions for the probability updating scheme defined by* (1) *to be EM are*

$$\frac{\theta_1(\mathbf{p})}{p_1} = \frac{\theta_2(\mathbf{p})}{p_2} = \cdots = \frac{\theta_R(\mathbf{p})}{p_R} = L \tag{2}$$

and

$$\theta_i(\mathbf{p}) - \phi_i(\mathbf{p}) = d_i.$$

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

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

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

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

The first two terms of the above involve the penalty probabilities, thus if $E[p_i(n+1)]$ is to be a linear function of $E[\mathbf{p}(n)]$, 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 *unknown* penalty probabilities $\{c_i\}$. Hence, *a* set of necessary and sufficient conditions for the scheme to be EM is

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

Consider the last two terms

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

We contend that these terms are linear in **p** if and only if

$$\frac{\theta_1}{p_1}=\frac{\theta_2}{p_2}=\cdots=\frac{\theta_R}{p_R}. \tag{3}$$

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

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

Hence (3) is obviously a sufficient constraint.

We prove necessity of (3) by considering the RHS as a linear function in **p**. In the most general case, for the last two terms to be linear in **p**,

$$p_j - p_j\sum_{i=1}^{R}\theta_i+\theta_j = \sum_{k=1}^{R}x_{j,k}p_k, \tag{5}$$

where $\{x_{j,k}\}$ is a set of nonnegative constants. Summing (5) over all values of j shows the LHS to be

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

which is unity.

The sum of the RHS of (5) over all values of j evaluates to

$$\sum_{j=1}^{R}X_jp_j, \qquad \text{where} \quad X_j=\sum_{k=1}^{R}x_{j,k},$$

which is unity if and only if every X_j is identically equal to unity. Hence (3) is necessary for the system to be EM, and the theorem is proved.

REMARK. The linear constraint involving $\phi(\cdot)$ and $\theta(\cdot)$ is

$$\phi_i(\mathbf{p}) - \theta_i(\mathbf{p}) = d_i.$$

Since the penalty probabilities are unknown, with no *a priori* information there is no loss in generality in 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

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

In matrix form, $E[\mathbf{p}(n+1)] = A^T E[\mathbf{p}(n)]$, where the elements of the stochastic matrix A are given by

$$A_{i,i} = 1-(1-d)c_i,$$

$$A_{j,i} = (1-d)\frac{c_j}{R-1}.$$

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* $\phi_i(\cdot)$ *and* $\theta_i(\cdot)$.

Proof. Subject to the conditions specified by Theorem I, the expected value of the action probabilities obeys the matrix 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 $\phi_i(\cdot)$ and $\theta_i(\cdot)$. Hence the theorem.

For the rest of this paper we shall assume that the d_i's are all equal. In any particular problem, if there is a reason to prefer one action over the other, the d_i's will be distinct. In such a case the matrix relationship $E[\mathbf{p}(n+1)] =$

$A^T E[\mathbf{p}(n)]$ will still be obeyed except that the matrix A will involve the set of parameters, $\{d_i\}$, as opposed to a *single* parameter d. It is our conjecture that by suitably choosing the d_i's, the *a priori* information about the actions can be included to render the scheme ε-optimal. This conjecture is currently being investigated. However, as stated earlier, for the rest of this paper, we shall assume that all the actions are initially equally preferred, and so the d_i's are all equal to a single constant, 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*

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

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

$$\mathbf{p}^* = A^T \mathbf{p}^*.$$

$\mathbf{p}^*$ is thus the eigenvector of the eigenvalue which is unity. Solving, $[I - A]^T \mathbf{p}^* = 0$ yields for the first row

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

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,

$$J = \frac{R}{\sum_{i=1}^{R} 1/c_i}.$$

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 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 greater than the arithmetic mean.

REMARKS.

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

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

$$\frac{\theta_j}{p_j} = a + bp_1 p_2 \cdots p_R = L.$$

When $b = 0$ and $d \neq a$, the scheme obtained is

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

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 I. 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, of which one *solely* controls the rate of convergence and the second, a, can be used to minimize the variance independently.

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 expediency. These conditions do not permit the linear dependence of $\phi(\cdot)$ and $\theta(\cdot)$, 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 (2) above. If the penalty probabilities are known (though the actions to which they belong are unknown), the design process is rendered easier. 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

$$c_1 = 0.65, \qquad c_2 = 0.2, \qquad c_3 = 0.5,$$

$$c_4 = 0.4, \qquad c_5 = 0.85.$$

Observe that a_2 is the optimal action and this action is chosen asymptotically with an expected probability

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

The nonlinear scheme which was used had the following form:

$$\frac{\theta_j}{p_j} = a + bp_1 p_2 p_3 p_4 p_5$$

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

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

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

as a function of b. In the above expression, N is the number of experiments,

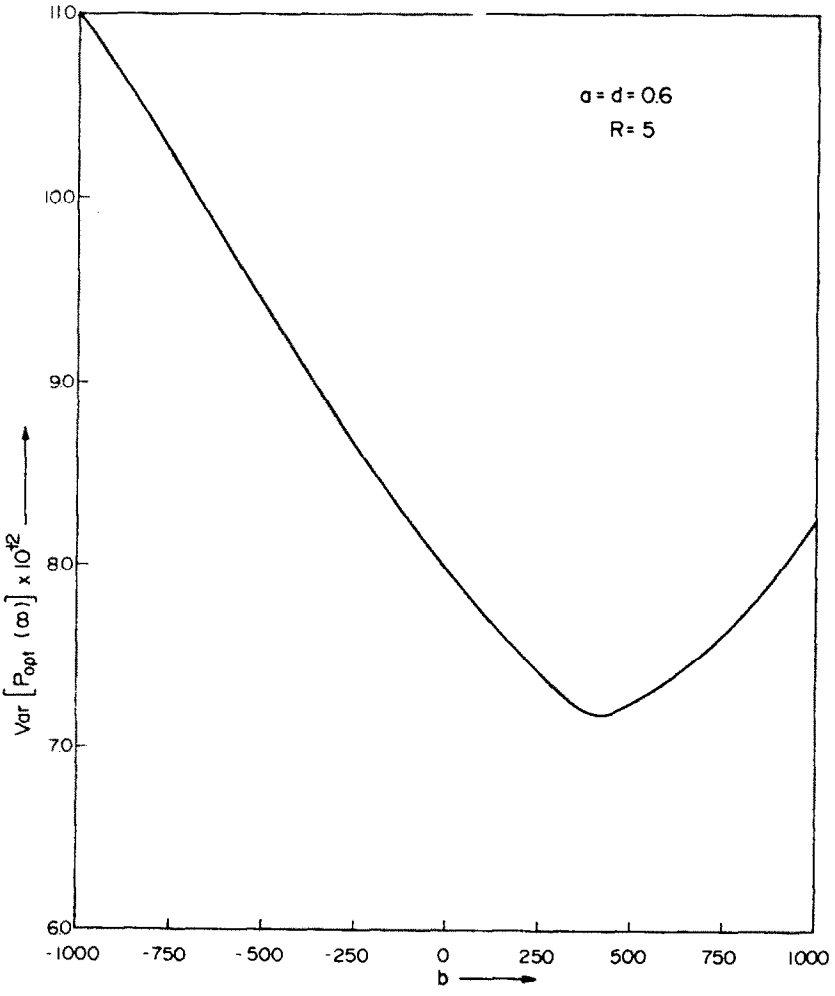

Fig. 2. Nonlinear EM scheme: variation of variance with b.

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 Figure 2 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. In contrast with the two-action EM schemes [11], however, due to the complexity of the expressions involved, 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).

We have considered the general problem of designing multiaction variable structure stochastic automata which are EM. The automata are fully defined by two families of probability updating functions $\phi_i(\cdot)$ and $\theta_i(\cdot)$. We have derived necessary and sufficient conditions on $\phi_i(\cdot)$ and $\theta_i(\cdot)$ that guarantee the scheme to be EM. The conditions on $\phi_i(\cdot)$ and $\theta_i(\cdot)$ 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 EM schemes.

In particular we have studied a whole family of linear schemes which are EM. Though these 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.

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. M. L. Tsetlin, On the behaviour of finite automata in random media, *Avtomat. i Telemekh.* 22:1345–1354 (1961).
2. M. L. Tsetlin, *Automaton Theory and the Modelling of Biological Systems*, Academic, New York, 1973.
3. A. Paz, *Introduction to Probabilistic Automata*, Academic, New York, 1971.
4. V. I. Varshavskii and I. P. Vorontsova, On the behaviour of stochastic automata with variable structure, *Avtomat. i Telemekh.* 24:327–333 (1963).
5. K. S. Narendra and M. A. L. Thathachar, to appear.
6. K. S. Narendra and M. A. L. Thathachar, Learning automata—a survey, *IEEE Trans. Systems Man Cybernet.* SMC-4:323–334 (1974).
7. D. L. Isaacson and R. W. Madson, *Markov Chains: Theory and Applications*, Wiley, 1976.
8. S. Lakshmivarahan and M. A. L. Thathachar, Absolutely expedient algorithms for stochastic automata, *IEEE Trans. Systems Man Cybernet.* SMC-3:281–286 (1973).
9. M. F. Norman, *Markov Processes and Learning Models*, Academic, New York, 1972.
10. M. F. Norman, Some convergence theorems for stochastic learning models with distance diminishing operators, *J. Math. Psych.* 5:61–101 (1968).
11. M. A. L. Thathachar and B. J. Oommen, Learning automata possessing ergodicity of the mean: The two action case, *IEEE Trans. Systems Man Cybernet.* Nov./Dec. 1983, pp. 1143–1148.
12. S. Lakishmivarahan, *Learning Algorithms Theory and Applications*, Springer, New York, 1981.
13. Y. A. Flerov, Some classes of multi-input automata, *J. Cybernet.* 2:112–122 (1972).
14. Y. Z. Tsypkin and A. S. Poznyak, Finite learning automata, *Engrg. Cybernetics* 10:478–490 (1972).
15. A. S. Poznyak, Use of learning automata for the control of random search, *Automat. Remote Control* 33(12):1992–2000 (1972).
16. Y. M. El-Fatteh, Gradient approach for recursive estimation and control in finite Markov chains, *Adv. in Appl. Probab.* 13:778–803 (1981).
17. Y. M. El-Fatteh, Multi-automaton games: A rationale for expedient collective behavior, *Systems Control Lett.* 1:332–339 (1982).

Received 12 December 1984; revised 8 March 1985