EPSILON-OPTIMAL STUBBORN LEARNING MECHANISMS+

J.P.R. Christensen* and B.J. Oommen**

SCR-TR-159, JUNE 1989

+ Partially supported by the Natural Sciences and Engineering Research Council of Canada and in part by the Copenhagen Telephone Company KTAS UB, 1199 København, Denmark.

* Copenhagen Telephone Company KTAS UB, 1199, København, Denmark. On leave of absence from Københavns Universitets Matematiske Institut, Universitetsparken, 2100, København, Denmark.

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

School of Computer Science, Carleton University
Ottawa, Canada, KIS 5B6

EPSILON-OPTIMAL STUBBORN LEARNING MECHANISMS+

J. P. R. Christensen* and B. J. Oommen**

ABSTRACT

In this paper we present a learning algorithm which has been but marginally referred to in the field of learning machines. The machine is an automaton whose structure changes with time and is assumed to be interacting with a random environment. The machine is essentially a stubborn machine. In other words, once the machine has chosen a particular action it **increases** the probability of choosing the action irrespective of whether the response from the environment was favourable or unfavourable. However this increase in the action probabilty is done in a systematic and methodical way so that the machine ultimately learns the best action which the environment offers. We show that the learning mechanism is ε-optimal and that the probability of it choosing the optimal action converges uniformly to unity. Apart from the fact that the machine is shown to be ε-optimal, a major contribution of this paper is that the mathematical tools used in the proof are quite novel to the field of learning. Besides the above theoretical results, the paper also contains various simulation results which demonstrate the properties of stubbornly learning mechanism. The mechanism is also shown to be inferior to the learning machine which merely ignores the penalty responses of the environment. Some open problems are also presented.

+ Partially supported by the Natural Sciences and Engineering Research Council of Canada and in part by the Copehnagen Telephone Company KTAS UB, 1199 København, Denmark.

* Copehnagen Telephone Company KTAS UB, 1199, København, Denmark. On leave of absence from Københavns Universitets Matematiske Institut, Universitetsparken, 2100 København, Denmark.

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

I. INTRODUCTION

The question of how animals and humans achieve learning has been the subject of study for a vast body of researchers. Initially, most of the research in this field was done by mathematical psychologists, and indeed, most of the pioneering results that are available are due to psychologists who had observed various learning trends in biological systems and who did their utmost to model the behaviour of these systems. The body of literature available for the modelling of psychological behavioral patterns is extensive [1-7, 9, 12-16, 21,22, 30,37] which, of course, includes the monumental references of Bush *et. al.* [3], Iosifescu *et. al.* [9], Krantz *et. al.* [10], Luce *et. al.* [14] and Norman[21]. Apart from the fact that these latter references describe the learning models for the behavioral patterns, they also describe in fair detail the the mathematical tools required to analytically study these models.

As opposed to mathematical psychologists, computer scientists and cybernetic engineers have, over the years, endeavoured to develop a perspective concept of learning. The differences between the two directions taken by psychologists and computer scientists has been portrayed by Lakshmivarahan in the introduction to his book[11]. Research that has emerged in the latter school of thought has been more in the direction of obtaining learning algorithms that satisfy various norms of learning and which can be implemented on a computer. Tsetlin [32] was probably the first researcher who proposed an automaton model for a learning system. Subsequently, various renowned researchers have been involved in developing pioneering results, and currently, a vast volume of literature is available in the field. Excellent review articles and books [11, 17, 18] are also currently available.

In the perspective model of learning, researchers have attempted to design learning machines which interact with an environment and which learn the optimal action which the environment offers. The machines are essentially stochastic automata whose structure is either fixed or time dependent. Automata of the latter class are called Variable Structure Stochastic Automata (VSSA). Learning automata have been used to tackle a variety of artificial intelligence problems including game playing, pattern recognition and hypothesis testing [17], priority assignment in a queueing system [11], telephone routing [11,17-19] and object partitioning [25].

Variable Structure Stochastic Automata (VSSA) can be completely defined by a set of action probability updating functions [11, 17, 35]. Once these updating rules are defined mathematically one observes that there is a vast overlap between the models proposed by the psychologists and the algorithms used by the researchers who have worked in the field from a perspective point of view. Indeed, some of the tools used to analyze the learning algorithms proposed use the foundational techniques presented by the earliest mathematical psychologists[3, 21].

Mathematical models have been available for a variety of psychological behavioral patterns. For example, the effect of reversal on pre-school children was studied by Campione [4], Dorfmann *et. al.* [5] presented a model for a continuum of sensory states, Estes *et. al.* [6] studied the effect of verbal conditioning, the overlearning reversal effect was analysed and studied by Lovejoy [13] and the role of attention in discrimination learning was studied by Zeaman [37].

In this paper we shall present a new learning model. There are organisms (yes, and even sometimes "rebellious" children) which seem to demonstrate a stubborn behaviour and yet be capable of achieving learning. As far as the mathematical psychologists are concerned, the model which we present represents a linear model which explains the psychological behaviour of these stubbornly learning organisms. In other words, even though they are penalized they seem to strengthen their opinion that their present choice is the best possible choice of action. However, if their learning reinforcement strategy is performed in a systematic fashion, we can prove that they can converge to the optimal action with a probability as close to unity as desired.

From the viewpoint of perspective learning the model which we present is indeed an algorithm. It is a linear algorithm with two parameters, and at every time instant, the algorithm increases the probability of choosing the action that it has just chosen. In other words once the machine has chosen a particular action it **increases** the probability of choosing that action **irrespective** of whether the response from the environment was a reward or a penalty. However, since the actions are chosen using an action probability vector (as opposed to being chosen using a fixed transition matrix as in the case of Tsetlin's automaton), it need not necessarily choose the same action at the next time instant. Thus, although the algorithm is stubborn in its myopic view, on

the long run, it has a way by which it can unlearn its "stubborn" choices and thus eventually arrive at the optimal action.

Throughout this paper we shall use the terms "model", "algorithm", "scheme" and "strategy" interchangeably. It should be observed however that the term "model" is suited only when the problem is viewed from the viewpoint of the descriptive mathematical psychologist. The perspective analyst would rather refer to the probability reinforcement mechanism as an algorithm, scheme or strategy.

The theoretical results which we present in the paper first proves that the scheme is absolutely expedient [11,17]. The main result of the paper is that the probability that the mechanism chooses the optimal action converges uniformly to unity, and thus that the strategy is ε-optimal. Indeed, we claim the powerful result that if the initial probability of choosing the optimal action is any value in the semi-closed interval (0,1], then the probability that it ultimately chooses the optimal action converges uniformly to unity. In practice however, since we work with real numbers of finite accuracy once the action probability of choosing the best action is fairly small, the automaton will tend to converge to the wrong action. This is primarily because, in practice, not only are {0, 1} the only set of absorbing states, but also, if the action probability assumes a value in the neighbourhood of these values, the likelihood of it moving away from the corresponding absorbing state is negligibly small.

Apart from the fact that the result is proved in its generality (i.e., specifying the general constraints on the parameters of the algorithm), one of the major contributions of this paper is that we have presented some mathematical tools which have, to the best of our knowledge, not been used to prove the ε-optimality of any schemes reported in the literature. These tools are some of the more advanced concepts of the theory of distributions and kernels [31]. The details of our *modus operandus* in proving the results and the difference between our techniques and the methods currently used in the literature will also be presented in the appropriate section.

Apart from the above theoretical results the paper also contains various simulation results which demonstrate the properties of the automaton discussed. The rate of convergence of this automaton in comparison with the traditional Linear Reward-Inaction (L_{RI}) scheme is also presented.

For the rest of this section we introduce the fundamentals and the notation which we shall use. Subsequently, we present the learning algorithm (model) and prove the various theoretical results we have obtained concerning its behaviour. We shall then present simulation results and compare the learning machine with the L_{RI} scheme.

I.1 Fundamentals

The learning mechanism presented in this paper is intended to represent one of two things. From the viewpoint of the descriptive analyst it represents a stubbornly learning organism which has be modelled. However, from the viewpoint of the perspective analyst it represents a learning algorithm or strategy. The learning mechanism interacts with a random environment (Figure I) and selects an action $\alpha(n)$ at each instant 'n' from a finite action set $\{\alpha_i \mid i = 1 \text{ to } R\}$. The selection is done on the basis of a probability distribution $\mathbf{P}(n)$, an R x 1 vector where, $\mathbf{P}(n) = [p_1(n), p_2(n), \ldots, p_R(n)]^T$ with,

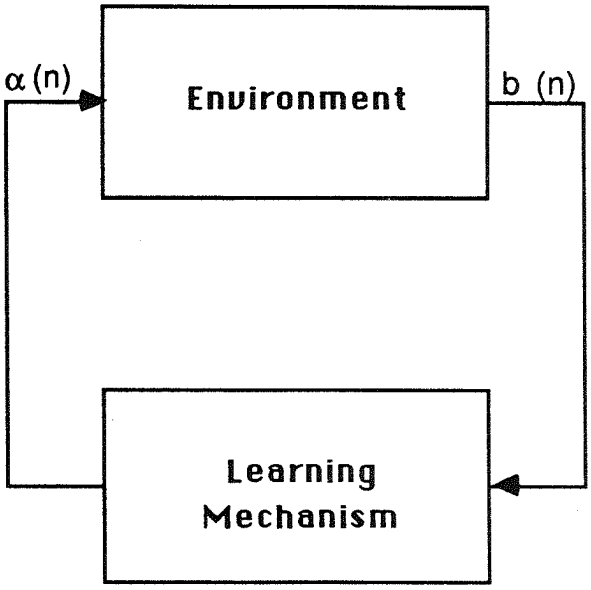
$$p_i(n) = \Pr[\alpha(n) = \alpha_i],$$

$$\sum_{i=1}^{R} p_i(n) = 1 \qquad \text{for all } n \tag{1}$$

The selected action serves as the input to the environment. Once the environment knows the action chosen by the mechanism it responds with a response b(n) at time 'n' . b(n) is an element of B = {0,1}. The response '1' is said to be a 'penalty'. The environment penalizes the mechanism with the time invariant penalty probability c_i, where,

$$c_i = \Pr[b(n) = 1 \mid \alpha(n) = \alpha_i] \qquad (i = 1 \text{ to } R). \tag{2}$$

Observe that the environment characteristics are completely specified by the set $\{c_i \mid i = 1 \text{ to } R\}$. On the basis of the response b(n) the action probability vector $\mathbf{P}(n)$ is updated and a new action chosen at (n+1).



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

$$\alpha(n) \in \{\alpha_1, \alpha_2, ..., \alpha_R\}$$

Figure I : The schematic of the interaction between the Environment and the Learning Mechanism

To render the problem nontrivial it is assumed that the set $\{c_i\}$ is unknown initially and it is desired that the mechanism arrives at the action which yields the minimum penalty as a result of interacting with the environment. The latter action, denoted as α_b, is referred to as the "best" action. It is assumed that the best action is unique and that the value of the penalty probability for this action is c_b. Observe that if the mechanism learns that this action is the best one it is equivalent to it choosing that action with a probability of unity and choosing all the other actions with a probability of zero.

I.2 Learning Criteria

Let us denote the expected penalty at time 'n' as E[M(n)] and the initial expected penalty to be M_0. Clearly, with no *a priori* information, since the learning mechanism chooses the actions with equal probability, M_0 will equal the arithmetic mean of the penalty probabilities. The mechanism is said to learn **expediently** if the limiting value

of E[M(n)] is less than M_0. The learning mechanism is said to be **optimal** if the limiting value of E[M(n)] equals the minimum penalty probability. It is **ε-optimal** if in the limit $E[M(n)] < c_b + \varepsilon$ for any arbitrary $\varepsilon > 0$. The closeness of the asymptotic value of E[M(n)] to c_b is generally achieved by suitably choosing some parameter of the automaton.

The mechanism is said to be optimal in probability if $p_b(n)$ asymptotically tends to unity in probability. It can be shown that this implies ε-optimality. Finally, the learning mechanism is said to learn in an **absolutely expedient** manner if at all time instances $E[M(n+1) \mid \mathbf{P}(n)] < M(n)$. Note that in this case M(n) is a supermartingale [8].

Throughout this paper we deal with the case when R, the numbers of actions, is two. The analogous results for $R > 2$ are yet open but conjectured to be true.

I.3 Note on Distribution Theory

The main results in this paper involve the ε-optimality of the stubbornly learning mechanism which we have proposed. However, as opposed to the traditional methods used in proving ε-optimality, in this paper, we have used as the basic tools some of the more advanced concepts of the theory of distributions and kernels. We refer the reader to an excellent treatise by Treves [31] for a detailed study on the latter theory. However, for the less informed reader, a short introduction into the field of distribution theory is not out of place.

For the purpose of this paper it is not sufficient to merely treat distributions in a purely intuitive way as "what you would get when you perform illegal differentiations". A distribution is to be thought of intuitively as a "function" and also formally to be denoted like a function. Any locally integrable function f on $\mathbb{I}^o$, the open unit interval, can be identified with a distribution when f is identified with a linear functional on the vector space of compactly supported infinitely differentiable functions on $\mathbb{I}^o$, the open unit interval. These functions are called "test functions". The distribution identified by f can be computed by integrating the product of f and a test function over $\mathbb{I}^o$ with respect to the normalized Lebesgue measure. The distribution on $\mathbb{I}^o$ and the space of test functions have weak and strong topologies and these are discussed in great detail by

Treves [31]. The fundamental properties which we will use in this paper are the strong strict inductive limit topology on the set of test functions of uniform convergence on all compact sets of all derivatives and the weak topology on the space of distributions.

Distributions are used in harmonic analysis and the theory of differential equations to vastly extend the domain of certain linear operators. In this paper we shall merely use this theory to identify a limit by showing that this limit satisfies a differential equation in the sense of distributions.

II. A LINEAR REWARD-REWARD LEARNING MECHANISM

When the field of learning automata was in its infancy, Lakshmivarahan and Thathachar [12, 17] proved a powerful result which presented necessary and sufficient conditions for a VSSA to be absolutely expedient. These conditions are quite all-encompassing and permit the probability updating functions to be fairly general functions of the action chosen. Indeed, this result was so all-encompassing that it included all the well known linear and nonlinear ε-optimal VSSA of their day. Aso and Kimura [2] generalized their result by describing a larger class of absolutely expedient automata in which the probability updating functions were a function of both the action chosen and of the action whose probability was to be updated. The learning mechanism which we present in this paper can be shown to be a member of the family described in [17].

With the above serving as a preamble we shall present the details of the stubbornly learning mechanism. The mechanism is a linear scheme which increases the probability of choosing the action that it has **just** chosen independent of whether the response from the environment was a reward or a penalty. We shall first formulate the mechanism and introduce it in the context of some historical remarks regarding its discovery. We shall proceed to show, using some elementary computations, that the scheme is absolutely expedient.

II.1 The Actual Learning Mechanism

The Linear Reward-Reward (L_{RR}) mechanism is explicitly defined for the two action case in terms of the probability updating rule defined for the probability $p_1(n)$ as follows :

$$\begin{aligned} p_1(n+1) &= ap_1 && \text{if } a(n) = \alpha_2 \text{ and } b(n) = 0 \\ &= 1 - ap_2 && \text{if } a(n) = \alpha_1 \text{ and } b(n) = 0 \\ &= (a + b)p_1 && \text{if } a(n) = \alpha_2 \text{ and } b(n) = 1 \\ &= 1 - (a + b)p_2 && \text{if } a(n) = \alpha_1 \text{ and } b(n) = 1 \end{aligned} \qquad (3)$$

where a, b >0 and (a+b) ≤ 1. In future, as in (3), for the sake of brevity, we shall omit the reference to the time instant 'n', and thus, with no loss of generality, p_i refers to the quantity $p_i(n)$. To render the problem meaningful we assume, with no loss of generality, that $p_1(0)$ is in the open unit interval. Also note that the scheme reduces to the traditional Linear Reward-Inaction (L_{RI}) scheme whenever a+b=1.

By symmetry, expressions analagous to (3) can be written down for the updating of $p_2(n)$.

It is clear that the linear scheme defined by (3) is of a Reward-Reward flavour whenever a+b < 1, because the mechanism systematically **increases** its probability of choosing α_i whenever α_i is chosen This mechanism was first reported by Viswanathan [36] in his doctoral thesis. It was later pointed out that this scheme can be also obtained as a special case of the schemes generated when the necessary and sufficient conditions required for absolutely expediency are marginally modified [10, 11, 17]. More recently, the second author of this paper [26] arrived at the same scheme by continuously changing the parameters of an ergodic VSSA which was capable of capturing the *a priori* information possessed by a learning automaton. In [26] the statement that this scheme fell within the category of the more general automata defined in [2] was erroneously made.

We now proceed to prove the theoretical properties of the mechanism. For the sake of fixing ideas, in what follows we shall assume that $c_1 < c_2$, and consequently that α_1 is the superior action.

II.2 Theoretical Properties of the Learning Mechanism

Theorem I

The L_{RR} scheme defined by (3) is absolutely expedient.

Proof :

By virtue of (3) observe that $p_1(n+1) - p_1(n)$ has the following distribution :

$$\begin{aligned} p_1(n+1) - p_1(n) &= -(1-a)\,p_1 && \text{with probability } p_2(1-c_2) \\ &= 1 - ap_2 - p_1 && \text{with probability } p_1(1-c_1) \\ &= -(1-a-b)\,p_1 && \text{with probability } p_2\,c_2 \\ &= 1 - (a+b)\,p_2 - p_1 && \text{with probability } p_1\,c_1 \end{aligned} \qquad (4)$$

But $1 - p_1 = p_2$. Hence $E[\,p_1(n+1) - p_1(n) \mid \mathbf{P}]$ has the form :

$$\begin{aligned} E[\,p_1(n+1) - p_1(n) \mid \mathbf{P}] &= -(1-a)p_1p_2(1-c_2) + (1-a)p_1p_2(1-c_1) - (1-a-b)p_1p_2c_2 \\ &\quad + (1-a-b)p_1p_2c_1 \\ &= bp_1p_2\,(c_2 - c_1). \end{aligned} \qquad (5)$$

Similarly, $E[\,p_2(n+1) - p_2(n) \mid \mathbf{P}] = bp_1p_2(c_1 - c_2)$. (6)

Let $\Delta M(n) = E[\,M(n+1) - M(n) \mid \mathbf{P}]$.

Then, since $M(n+1) = c_1\,p_1(n+1) + c_2\,p_2(n+1)$,

$$\Delta M(n) = c_1 E[\,p_1(n+1) - p_1(n) \mid \mathbf{P}] + c_2\,E[\,p_2(n+1) - p_2(n) \mid \mathbf{P}].$$

Substituting (4) and (5) we obtain,

$$\begin{aligned} \Delta M(n) &= c_1bp_1p_2\,(c_2 - c_1) + c_2\,bp_1p_2\,(c_1 - c_2\,) \\ &= -bp_1p_2\,(c_2 - c_1)^2 \;<\; 0. \end{aligned}$$

Hence, M(n) is a supermartingale and the theorem is proved. •••

As explained in the preamble to this section, although the L_{RR} scheme does not, strictly speaking, fall into the class of absolutely expedient schemes described in [12], it

can be seen to satisfy analogous symmetry conditions as those defined in [12] by appropriately marginally modifying the upper and lower bounds on the penalty functions[17]. Indeed, using the notation of [12], the L_{RR} scheme can be put in the form :

$\phi_j(\mathbf{p}) = (1\text{-}a).p_j$, and,

$\psi_j(\mathbf{p}) = -(1\text{-}a\text{-}b).p_j$.

Hence, for all j,

$\phi_j / p_j = (1\text{-}a) = \lambda$, and, $\psi_j / p_j = -(1\text{-}a\text{-}b) = \mu$.

Thus, the symmetry conditions of [12] are satisfied, but μ has a negative sign. This has been clarified in greater detail in [17].

It is also interesting to note that the learning mechanism has to pay for its stubbornness. We shall show that although the scheme is ε-optimal, the fact that on being penalized it increases the action probability of the action that it has just chosen, is not to its advantage. Indeed, it would have been much better if it had just ignored the penalty response and left the action probabilities unchanged. This was shown in [26], but for the sake of completeness it is repeated below. It is well known that for the L_{RI} scheme,

$$\Delta M(n) = -(1\text{-}a)p_1p_2 (c_2 - c_1)^2 \ < \ 0.$$

However, for the L_{RR} scheme,

$$\Delta M(n) = \ -bp_1p_2 (c_2 - c_1)^2 \ < \ 0.$$

In the latter case, since (1-a-b) > 0 has to satisfied for maintaining the Reward-Reward character of the scheme the above two inequalities lead to :

$$|\Delta M(n)|_{RI} \geq |\Delta M(n)|_{RR}$$

Thus the L_{RI} scheme is superior to the L_{RR} scheme because the single-step decrement in the mean penalty is greater for L_{RI} scheme than for the L_{RR} scheme. Thus the learning mechanism has to "pay" for its stubborness because it would have been much better off if it had ignored the penalty response rather than stubbornly increasing the probability of choosing the action that it just chose. This is reminiscent of the saying "One has to sleep on the bed that he himself has made".

Apart from being absolutely expedient we shall now show using distribution theory that that the scheme is ε-optimal. In other words, we prove that if $c_1 < c_2$, the probability that the automaton converges into the absorbing barrier $[1, 0]^T$ can be made as close to unity as desired by suitably choosing the parameters of the scheme, provided that the initial probability of choosing α_1 is not in the neighbourhood of zero. The exact formulation of the latter constraint will be made more specific presently.

Before we embark on the actual proofs of the theorems, we introduce the following definitions.

Definitions

(i) Lower Semi-continuous Functions

A real valued function f on a topological space $\mathfrak{X}$ is called lower semi-continuous if and only if for any real number $r \in \mathfrak{R}$, the subset of f defined by

$$\{ x \in \mathfrak{X} \mid f(x) > r \}$$

is open in $\mathfrak{X}$.

Note that any supremum of any family of lower semi-continuous functions is again lower semi-continuous because the union of open sets is open. This is the reason why lower semi-continuous functions are used so extensively in analysis. If a lower semi-continuous function f on the real line $\mathfrak{R}$ is increasing, then for every $x \in \mathfrak{R}$, the limit from the left, f(x-0), equals f(x), and the number of points where there are discontinuities from the right (i.e., where $f(x+0) \neq f(x)$) is at most countable.

(ii) The Markov Chain $\mathfrak{M}(x)$

Let the random variable x_n represent the action probability $p_1(n)$. Note that if $\mathfrak{I}=[0,1]$, $x_n \in \mathfrak{I}$. Also let $\mathfrak{M}(x)$ represent the Markov chain $\{x_n \mid n \geq 0\}$.

Observe that the set of equations represented by (4) completely defines $\mathfrak{M}(x)$. We now prove the properties of $\mathfrak{M}(x)$.

Theorem II

The chain $\mathfrak{M}(x)$ converges almost surely either to 0 or to 1.

Proof :

Using the conditional expectation of x_{n+1} given x_n from (5) and (6) we have,

$$E[x_{n+1} \mid x_n] = x_n + b(c_2-c_1)(1-x_n)x_n.$$

Hence the sequence $\{x_n\}$ of random variables is a submartingale and since the sequence is uniformly bounded, it has a limit, L, almost surely. The expectation $E[x_n]$ is guaranteed to not tend to infinity if and only if this limit L satisfies:

$$b(c_2-c_1)(1-L)L = 0 \text{ almost surely.}$$

This proves the Theorem. •••

The reader will easily associate the above theorem as an application of the well-acclaimed martingale convergence theorem [20]. The theorem has been included here to give the reader an insight into the ramifications of the result in the context of our learning scheme.

Since the scheme converges almost surely to the values in {0,1} we see that our problem is conceptually simplified into one of computing the probability of it converging to unity. Clearly, this probability is a function of the initial probability x_0. To aid in the computation of the probability of absorption into the stae {1} we define f(x) as the probability that the Markov chain $\mathfrak{M}(x)$ tends to 1. Of course f(x) is defined on $\mathfrak{I}$ and has values in $\mathfrak{I}$. From the definition of f(x) it follows that f(·) satisfies the following functional equation :

$$(Tf)(x) = (1-x)(1-c_2)f(ax) + (1-x)c_2f((a+b)x)$$
$$+ x(1-c_1)f(1-a(1-x)) + xc_1f(1-(a+b)(1-x)) = f(x). \qquad (7)$$

Although this equation is complicated we shall use it to study of the limiting behaviour of f(x) under the above limiting operation.

Theorem III

The function f defined above satisfies $x \leq f(x) \leq 1$, and is both increasing and lower semi-continuous. Furthermore f is continuous on the half open interval (0,1].

Proof :

By virtue of (3), the linear operator T is defined by :

$$(Tg)(x) = (1-x)(1-c_2)g(ax) + (1-x)c_2 g((a+b)x) + x(1-c_1)g(1-a(1-x)) + xc_1 g(1-(a+b)(1-x)). \quad (8)$$

Clearly T maps the linear space of real functions on $\mathfrak{I}$ into itself. If g is any continuous function on $\mathfrak{I}$ and we apply T^n to g, then as n tends to infinity, $T^n(g)$ tends pointwise to the function $g(0)(1-f(x)) + g(1)f(x)$ (because of the above discussed limiting behaviour of the Markov chain $\mathfrak{M}(x)$). This, of course, implies that the only continous solutions to the functional equation (7) are sums of multiples of f and a constant. From this fact, however, it does not follow, that f is continous, but we shall tackle this question later.

A special case of the above remark is that if $g(x)=x$, T^n applied to g(x) converges pointwise to f(x). This follows since, in this case,

$$(Tg)(x) = x + b(c_2-c_1)\, x\, (1-x)$$

and by induction it is easy to see that indeed $T^n(g)$ tends monotonously to f. Hence f is lower semi-continous.

To show that f is increasing, we now show that $T^n(g)$ is an increasing function, and this is achieved by induction on n. Consider the definition of the operator T, as per (8). Let g be a continously differentiable function. We shall show, that Tg is increasing if g is increasing. To do this we shall apply the differential operator D to Tg and show that the resulting function is non-negative.

The function Tg is a sum of four terms and D(Tg) thus is a sum of eight terms, since the terms in Tg are products. These terms are :

(i) $(1-c_2)\, g(ax)$

(ii) $a(1-x)(1-c_2)\, Dg(ax)$

(iii) $(-c_2)\, g((a+b)x)$

(iv) $(a+b)(1-x)c_2\, Dg((a+b)x)$

(v) $(1-c_1)\, g(1-a(1-x))$

(vi) $ax(1-c_1)\, Dg(1-a(1-x))$

(vii) $c_1\, g(1-(a+b)(1-x))$

(viii) $(a+b)xc_1\, Dg(1-(a+b)(1-x))$

Note that of the above, the first and the third terms are non-positive. Consider the sum of the fifth and seventh terms. By some simple manipulations it is easy to see that the sum of these two terms satisfies :

$$(1-c_1)g(1-a(1-x)) + c_1 g(1-(a+b)(1-x)) \quad > \quad g(1-(a+b)(1-x)) \quad > \quad g((a+b)x). \quad (9)$$

In the above we have (repeatedly) implicitly made use of the assumption that g is increasing. Collecting the first and third terms with the above two positive terms and comparing the expressions algebraically it follows that the sum of these four terms is non-negative. Furthermore, the sum of the other four terms is also non-negative since they individually are non-negative. Thus, D(Tg) is non-negative. Thus the first part of Theorem II follows from the fact that f is a limit of a monotonously increasing sequence of continous increasing functions.

To show the continuity of f(x), we note that because of the functional equation (7), a discrete jump in a value of x in $\mathfrak{I}^0$ would cause a jump in f(x) of at least the same size in at least one of the four points to which $\mathfrak{M}(x)$ may move. Thus, by induction, we can see that there exists an arbitrary large number of jumps of at least this size. This is a contradiction and the theorem is proved. •••

We note in passing that it is as yet not clear to us whether f needs to be continous at zero. However, we are now ready to formulate and prove the main results of the paper.

Theorem IV

For the scheme defined by (3), let $a \to 1$, $b \to 0$ and $b > \delta(1-a)$. Also, if T is defined as in (8) and $h=(1-a)/a$, then, $((Tf)(x)-f(x))$ can be written as a sum of four difference operators P_1, P_2, P_3 and P_4, where,

(i) $(1/h)P_1 f$ tends weakly to $-x(1-x)(1-c_2)Df$ as $h \to 0$,

(ii) $(1/h)P_2 f$ tends weakly to the distribution $-\sigma x(1-x)c_2 Df$ as $h \to 0$,

(iii) $(1/h)P_3 f$ tends weakly to the distribution $x(1-x)(1-c_1)Df$ as $h \to 0$

(iv) $(1/h)P_4 f$ tends weakly to the distribution $\sigma x(1-x)c_1 Df$ as $h \to 0$

where Df is to understood in the sense of distributions.

Proof :

Consider a slight transformation of (7) to evaluate $((Tf)(x)-f(x))$. Indeed, after some manipulation we see that,

$$((Tf)(x)-f(x)) = (1-x)(1-c_2)(f(ax)-f(x)) + (1-x)c_2(f((a+b)x)-f(x)) + x(1-c_1)(f(1-a(1-x))-f(x)) + xc_1(f(1-(a+b)(1-x))-f(x)) \quad (10)$$

By using difference operators (10) can be equivalently written as :

$$(P f)(x) = (P_1 f)(x) + (P_2 f)(x) + (P_3 f)(x) + (P_4 f)(x) \quad (10b)$$

where each P_i is a difference operator and each $(P_i f)(x)$ represents the corresponding term in (10) above.

Under the limit operation we now perform the following substitutions :

$a = 1/(1+h)$, and,

$a+b = 1/(1+k)$.

The limit $a \to 1$ now monotonously translates to the limit $h \to 0$. Also, the corresponding limit for k is $k \to 0$, although this translation is not monotonous. Also, the condition $b > \delta(1-a)$ translates to the equivalent condition $k/h < (1-\delta)$. We shall now let $h \to 0$ and $k \to 0$ subject to the constraints of the theorem, namely that $b > \delta(1-a)$ (i.e., that $k/h < (1-\delta)$).

We intend to study the limit behaviour of the difference operator $(1/h)P$ as h tends to zero.The topology in which this limit operation is achieved will be discussed

presently. Notice that a potential problem could be encountered in computing the limit since the function f depends both on h and k. We shall overcome this difficulty later on, but for the time being we consider f as a fixed function and let h tend to zero. We shall show that each of the four terms of this operator tends to be a differential operator as h tends to zero, and thus in the limit (1/h)P also tends to be a differential operator.

Let ϕ be any infinitely differentiable real function on $\mathfrak{I}^o$ with a compact support in the open unit interval. As a computational convenience we introduce the function $\theta(x)$ which is assigned to have the value $(1-x)(1-c_2)\phi(x)$. Clearly, θ is an infinitely differentiable real function with compact support contained in $\mathfrak{I}^o$. For a moment let us assume that f is not changed when $h \to 0$. Under these circumstances we shall study the limiting behaviour of the expression :

$$(1/h) \int_0^1 \{ f(ax) - f(x) \} \theta(x) \, dx \qquad (11)$$

We shall first endeavour to explain the role and need of this integral in our proof. The above integral is precisely the inner product in the distribution sense of the test function ϕ with the difference operator $(1/h)P_1$ applied to the function f, where P_1 is defined in (10b) and f itself is considered as a distribution. Now if we temporarily overlook the fact that f changes under the limit, and consider f to be a fixed distribution, then the discussion below implies that $(1/h)$ $(P_1 f)$ tends weakly to the distribution $-x(1-x)(1-c_2)$ $Df(x)$, or that the difference operator $(1/h)P_1$ tends weakly to $-x(1-x)(1-c_2)D$. In an analogous manner a similar evaluation of all the four operators P_1 P_2, P_3, and P_4 is acheived and the resulting equation represented by (10b) is shown to be a differential equation which has a constant solution.

The integral (11) consists of two terms. The second term is not changed at all in the limiting process. Let us consider the value of the first term of the integral. We simplify the first portion of the integrand by performing the substitution of x by a dummy variable y where :

$x = (1/a)y = (1+h)y.$

Observe that by performing a simple change of variables and the subsequent resubstitution of y by x, the first term yields the integrand to be $f(x)\theta((1+h)x)(1+h)$, and the lower and upper limits of the integral would be 0 and $1/(1+h)$ respectively. But this upper limit may just as well be replaced by unity, when h is sufficiently small, since θ has a compact support in $\mathfrak{I}^o$. Thus, collecting the two terms and using the fact that the difference quotient for θ is uniformly bounded and tends uniformly to the differential quotient $D\theta$, we obtain the result that in the limit this integral is equal to the value of the integral of $f(x)(\theta(x)+xD\theta(x))$ between the lower limit of 0 and the upper limit 1.

To evaluate the latter integral we shall use the test function $\theta(x) := (1-x)(1-c_2)\phi(x)$, where $\phi(x)$ is an infinitely differentiable real function with a compact support contained in $\mathfrak{I}^o$.

Consider the integral of $f(x)(\theta(x)+xD\theta(x))$ between the limits 0 and 1. The latter integral is the inner product in the distribution sense between f (considered as a distribution) and (Id + xD) applied to θ, where Id represents the identity operator. Transposing the operators of multiplication and differentiation and observing that the multiplication operator is formally self-transposed and that the formal transpose of D is -D (see pp. 249 of Treves [31] for the definition of the formal transpose of a differential operator), we get the result that the integral is the inner product between the distribution given by the expression

$$\begin{aligned} f(x) + Dt(xf(x)) &= f(x) - f(x) - xDf(x) \\ &= - xDf(x) \end{aligned}$$

and the test function θ.

Resubstituting now the value of $\theta(x) = (1-x)(1-c_2)\phi(x)$ gives us the result that this is the inner product of $-x(1-x)(1-c_2)Df(x)$ and the test function ϕ, and thus, the value of the integral is the same as the value of integrating the function $-x(1-x)(1-c_2)Df(x)\phi(x)$ between the limits 0 and 1. Although for the time being we note that this is a formal computation, we shall later justify this differentiation process as a process which is to

be understood in the sense of distribution theory [31].

But this evaluation has demonstrated that if f is an L^1 function on $\mathbb{I}$ with respect to the Lebesgue measure, then $(1/h)P_1$ f tends weakly to $-x(1-x)(1-c_2)Df$ where Df is to understood in the sense of distributions, where, the weak convergence is the usual weak topology on the space of distributions on $\mathbb{I}^o$ induced by inner products with test functions.

Notice that all along we have worked with the first terrm of (10). We shall now proceed to perform analogous computations for the second, third and fourth terms in (10). The second term of (10) is now treated analogously and we get :

$$(1/h) \int_0^1 (f((a+b)x) - f(x))\, \theta(x)\, dx$$

$$= (k/h)\,(1/k) \int_0^1 (f((a+b)x) - f(x))\, \theta(x)\, dx \qquad (12)$$

We assume now, that the limit $h \to 0$, $k \to 0$, is taken in such a way that the quotient k/h has a limit σ which, of course, has to be in the interval $[0, 1-\delta]$. Under the latter assumption, just as we did for the first term of (10), we multiply the second term of (10) with the infinitely differentiable function ϕ having a compact support in $\mathbb{I}^o$, and integrate the expression with the limits from 0 to 1 and multiply with (1/h). Observe that if we substitute $(1-x)c_2\phi(x)$ by $\theta(x)$ the preceding discussions apply for this new test function θ. Again, to simplify the relevant portion of the integrand of (12) for this term we perform the substitution of x by a dummy variable y where, in this case,

$$x = (1/(a+b))y = (1+k)y.$$

By performing this simple change of variables and then subsequently resubstituting y by x, and finally grouping the terms together we see that the value of the limit is the value of the integral of $-\sigma x(1-x)c_2 Df(x)\phi(x)$ between 0 and 1. Thus $(1/h)P_2 f$ tends weakly to the distribution $-\sigma x(1-x)c_2 Df$ as $h \to 0$.

We now proceed in an analogous way to work with the third and fourth terms of (10). The substitution used for the dummy variable y is almost the same. In the case of the third term of (10) we use the substitution

$y = ax+(1-a)$ and $x = (1+h)y-h$

and at the end of the computations we resubstitute y by x. This time we get a very similar result, namely that the result is the integral of $f(x)(\theta(x) + (x-1)D\theta(x))$ integrated between the limits 0 and 1 which reduces to the integral of $x(1-x)(1-c_1)Df(x)\phi(x)$ between 0 and 1. Thus $(1/h)P_3f$ tends weakly to the distribution $x(1-x)(1-c_1)Df$ as $h\rightarrow 0$.

Finally, for the fourth term of (10) the substitution that is used introduces the dummy variable y, where,

$y = (a+b)x+(1-(a+b))$ and $x = (1+k)y-k$.

Again, at the end of the computations we resubstitute y by x to obtain the result that for the fourth term of (10) the final result is the integral of $\sigma x(1-x)c_1Df(x)\phi(x)$ between 0 and 1. Thus $(1/h)P_4f$ tends weakly to the distribution $\sigma x(1-x)c_1Df$ as $h\rightarrow 0$.

Hence the theorem. •••

Observe that the basic difference between the final results obtained for the operators P_1 and P_3 is the negative sign and the interchange of the coefficients c_1 and c_2. Similarly, the basic difference between the final results obtained for P_2 and P_4 is the negative sign and the interchange of the coefficients c_1 and c_2.

Using the above theorem we shall now formulate and prove our main result .

Theorem V

Let $z \in (0,1)$. Then for all $\delta \in (0,1)$ the probability that f(x) tends to 1 converges uniformly to 1 on the interval [z, 1] whenever $a\rightarrow 1$, $b\rightarrow 0$ and $b > \delta(1-a)$.

Proof :

Before embarking on the involved proof, we note that if b is just a constant equal to (1-a), the scheme reduces to the L_{RI} scheme, which is well known to be ε-optimal as

$a \rightarrow 1$. Thus, in the context of the learning mechanism under study, the theorem is meaningful only if b is not equal to (1-a). We shall show that the probability that f(x) tends to 1 converges uniformly to 1 as $a \rightarrow 1$ with the additional constraint that $b \rightarrow 0$ subject to $b > \delta(1-a)$. These are indeed the constraints of Theorem IV too !

The basic idea in the proof is that the equation defined by (10) above becomes a differential equation satisfied in the limit only by constants. And this result is also true if in the distribution sense [31]. The complications of the proof are due to the fact that both the equation and f(x) vary under the limit operation.

Consider (10) which evaluates ((Tf)(x)-f(x)). By performing some straightforward simplifications we **re-write** (10) to yield :

$$\begin{aligned}((Tf)(x)-f(x)) &= (1-x)(1-c_2)(f(ax)-f(x)) + (1-x)c_2(f((a+b)x)-f(x)) \\ &\quad + x(1-c_1)(f(1-a(1-x))-f(x)) + xc_1(f(1-(a+b)(1-x))-f(x)) \\ &= 0. \end{aligned} \tag{10}$$

Since the right hand side of (10) is identically equal to zero, we **re-write** (10b) as :

$$(P\ f)(x) = (P_1\ f)\ (x) + (P_2\ f)\ (x) + (P_3\ f)\ (x) + (P_4\ f)\ (x) = 0 \tag{10b}$$

where each P_i is a difference operator and each $(P_i\ f)\ (x)$ represents the corresponding term in (10) above.

By virtue of Theorem IV we have shown that if we multiply (10) with (1/h) and let $h \rightarrow 0$, (10) approaches a differential equation. Indeed, we shall show that this equation has only a constant solution, and subsequently deduce our result.

Adding together the four terms of (10b) using the differential operators derived in Theorem IV, we get that, independent of the distribution f on $\mathfrak{I}^o$, (1/h)P f(x) tends weakly to the distribution S f(x), where,

$$S\ f(x) = (-\ (1-c_2) - \sigma c_2 + (1-c_1) + \sigma c_1)x(1-x)\ Df(x). \tag{13}$$

The proof is now **informally** concluded as follows. The sum of the four terms is always equal to zero independent of what the test function ϕ happens to be. But this expresses the fact that the sum of the four terms given by (13) (understood in the distribution sense) vanishes. But since the coefficients do not vanish, this implies that f

is a constant function. We shall proceed to formulate the above arguments formally.

Since the family of functions f defined by Theorem III is a uniformly bounded family of measurable and even lower semi-continuous functions, we may assume that by letting $h \to 0$ through a universal net (and $k \to 0$ subject to the constraints stated above) that f has a certain weak limit point F in $L^2(u)$, where u is the normalized Lebesgue measure in the unit interval. Observe that we could alternatively have worked with sequences and subsequences. All computations referred to above are now immediately justified in the distribution theoretic sense [31] even if they are not in the standard sense in which we have a fixed real function ϕ, which is infinitely differentiable and which has a compact support in $\mathfrak{I}^o$. Indeed, we shall now view all the above computations in the distribution sense [31].

We want to show that, since the right hand sides of both (10) and (10b) evaluate to zero, using the notation of (13), F satisfies the differential equation, $SF(x) = 0$ in the sense of distribution theory [31]. Indeed, this will be obtained as a formal limit using the above computations. To see that this is true we note that our computations may be expressed by stating that the integral of f(x) multiplied by $(P^t\phi)(x)$ between the limits 0 and 1 is identically equal to zero, where, P^t is the transpose of the operator P. Note that this formal transpose operator can be evaluated to be the transpose of (1/h) multipled by the transpose of P defined by (10b). Here f varies as a function of h and k, as does P^t, and the latter also being a difference operator. The function ϕ, however, is assumed to be fixed. We are now in a position to justify the limit operations involved.

It is essential for our proof, to note that the family of functions $\{ P^t\phi\}$ (when $0 < d \leq h < 1$, d is fixed and the parameter k belongs to the interval (0,h]) is conditionally compact in the natural topology on the space of test functions. In other words, we consider the topology of uniform convergence on compact subsets of $\mathfrak{I}^o$ for derivates of any order. For the purpose of our proof, we only need to consider the topology of uniform convergence on compact sets for derivatives of order 0 (i.e., the function itself) and of order 1. If in the limiting cases, we interpret P^t as a differential operator, then indeed, the above statements hold for compact sets as well as for conditionally

compact subsets. Although the latter claim seems fairly involved to verify, it follows easily from the fact that with respect to this topology on the space of test functions, the difference operators applied to ϕ depends continously on the difference and indeed converges, in this topology, to the differential quotient as the difference tends to zero.

Let $H = \{f \in L^2(u) \mid x \leq f(x) \leq 1\}$.

Then the family of functions in $L^2(u)$ (u is the normalized Lebesgue measure on the unit interval) given by H is weakly compact in $L^2(u)$. From this we see that the inner product defined on the product of these two families is a jointly continuous function with respect to the product topology on the product space. This means that any sequence $h_n \in H$ for which $h_n \rightarrow F$ tends uniformly to F when the elements of H are considered as functions of $\{P^t\phi\}$. This enables us to go to the limit in a formally correct manner even though P, P^t and f vary under the limit. Thus, in the limit, f tends weakly to F, and therefore uniformly to F, where F is considered as a function on the family $\{P^t\phi\}$ via the natural inner product. Hence, in the limit, $P^t\phi$ tends to $S^t\phi$ in the topology of test functions, where, the operator S^t is the formal transpose of the operator S defined in (13). The joint continuity thus implies that the inner product of F with $S^t\phi$ is zero, and this further means that, in the sense of distributions [31], the inner product of S(F) with ϕ is zero. But since ϕ is an arbitrary smooth function with support in $\mathbb{I}^o$, we conclude that F satisfies S(F) = 0 in the sense of distribution theory [31]. Hence, F is a constant almost surely on the unit interval $\mathbb{I}$, and since we also have the inequalities on x as

$x \leq F(x) \leq 1$ almost surely

this constant must be unity.

Since in the limit, F is the constant function 1 whatever the universal subnet or subsequence used, we may conclude, that f tends weakly to 1 in the sense of distributions [31]. Because of the boundedness assumptions on f, we can therefore conclude that f tends weakly to 1 in $L^2(u)$. But this, in conjunction with the result of Theorem III which states that f is increasing formally asserts the uniformity statement of Theorem V. Hence the Theorem ! ●●●

Corollary V.1

The L_{RR} learning mechanism is ε-optimal.

Proof

The result is a direct consequence of Theorem V. •••

II.3 Remarks

(i) One of the primary contributions of this paper is the formulation and the study of a model for a stubborn learning mechanism. However, as stated in the introduction, such a learning mechanism was not entirely unknown to the literature [17, 26, 36]. To the best of our knowledge, it was Viswanathan [36] who first reported the scheme as a learning algorithm and alluded to its absolute expediency. The details of how the algorithm satisfies the symmetry conditions required for absolute expediency has been explained in detail in [17]. The fact that this scheme naturally falls out as a general case of an ergodic automaton capable of capturing *a priori* information was described in detail in [26].

(ii) The main contribution of this paper is not merely the formulation and the recognition of the stubborn learning mechanism. We believe that the fundamental contribution of this paper is the *modus operandus* whereby we prove the ε-optimality of the L_{RR} learning mechanism. In the literature the technique by which absolutely expedient automata have been proven to be ε-optimal is as follows. After deriving a functional equation analogous to (7) the existing techniques have attempted to formally solve the equation by obtaining increasingly accurate approximations of the solution. Indeed, it can be shown that the resultant solution is bounded from above by a superregular function and that it is bounded from below by a subregular function [11, 17]. The actual closed form expressions for these functions are then specified in terms of normalized exponentials whose coefficients can be computed based on the approximating function. Subsequently, by approximating the function that bounds the solution from below using the first few terms of its infinite series, it can be shown that if the parameters of the absolutely expedient learning algorithm are "small enough" the accuracy of the convergence of the scheme can be made as close to unity as desired.

As opposed to this approach, we have shown the ε-optimality of the L_{RR} learning mechanism in a completely distinct way. First of all, observe that we have proven the ε-optimal of the scheme even when the automaton is strictly of a Reward-Reward flavour. Secondly, we have not tried to approximate f(x) from below by a subregular function and then applied limiting arguments on this function as the parameters of the scheme are made correspondingly smaller (or larger, as would be the case of our parameter, 'a'). Notice that although both these are extremely elegant and ingenious operations, the actual process of obtaining these limiting coefficients can be quite cumbersome. However, in this case, we have repeatedly used the fact that f(x) itself is a distribution. Once this is done, the various salient features of the theory of distributions, kernels and topological spaces [31] have been used to arrive at the result that the probability that f(x) tends to 1 converges uniformly to unity as the limiting operation is achieved. To the best of our knowledge, such a method of reasoning is quite novel in the area of both descriptive and perspective learning theory. Also, as opposed to traditional methods, apart from showing that the scheme is ε-optimal when the initial action probability of choosing α_1 is 0.5, we prove that it is ε-optimal whenever the initial action probability of choosing $\alpha_1$is not in some specified neighbourhood of zero.

(iii) It is easy to see that a similar set of arguments can be given to prove that the L_{RI} scheme is ε-optimal. Notice that this result would be far more general than the result which exists in the literature [11, 17], because, as explained in the earlier paragraph, we assert that the scheme is ε-optimal even though the initial action probability of choosing α_1 is not 0.5. Indeed, we can prove that the scheme is ε-optimal whenever the initial action probability of choosing $\alpha_1$is not in some specified neighbourhood of zero.

III. SIMULATION RESULTS

To test the learning capability of the L_{RR} learning mechanism the latter has been simulated for various environments in which $c_2 = 0.8$ and c_1 was varied from 0.1 to 0.6. In each experiment the initial starting probability was always $[0.5, 0.5]^T$ and in each case the parameters of the scheme were varied to study the accuracy obtainable by the scheme . To obtain dependable results, each simulation was performed **400** times,

and the ensemble averages of these experiments is reported below.

In Table I, the variation of $\hat{E}\ [\ p_1(\infty)\ ]$ is presented when $c_2 = 0.8$ and c_1 is varied from 0.1 to 0.6. In all the cases, 'a' was set equal to 0.9 and 'b' was varied from 0.01 to 0.09. In Table I we have reported the results obtained for 'b' having the value 0.09. In this case, in every single experiment the scheme converged to the correct action, and indeed, this was true for all the values of c_1 even as it varied from 0.1 to 0.6.

To compare the rate of convergence of the scheme with the L_{RI} scheme (obtained when a+b is unity) the actual mean time to converge to 99% of the final action probability has also been recorded in Table I. Thus, when c_1 is 0.1 and c_2 is 0.8, the mean time for the L_{RR} scheme to converge to an accuracy of 99 % was 79 iterations. The corresponding mean time to converge for the L_{RI} scheme was 70 iterations. The corresponding figures for c_1=0.6 and c_2 =0.8 are 333 for the L_{RR} scheme and 257 for the L_{RI} scheme respectively. Thus, the latter scheme seems to be categorically faster than the L_{RR}, just as was predicted by the derivation following Theorem I.

c_2	'b'	$E[p_1(\infty)]$	M.T.C.
0.1	0.09	1.00	79
	0.1	1.00	70
0.2	0.09	1.00	93
	0.1	1.00	83
0.3	0.09	1.00	112
	0.1	1.00	102
0.4	0.09	1.00	140
	0.1	1.00	127
0.5	0.09	1.00	200
	0.1	1.00	171
0.6	0.09	1.00	333
	0.1	1.00	257

Table I: Simulation results involving the L_{RR} learning mechanism. In each case a=0.9. The case when a=0.9 and b=0.1 represents the L_{RI} scheme.

IV. CONCLUSIONS

In this paper, we have considered the problem of a learning mechanism learning the optimal action offered by a random environment. The mechanism which we have presented can be defined as an action probability updating rule and thus from the viewpoint of perspective theoretician, it is a Variable Structure Stochastic Automaton (VSSA). The machine is essentially a stubborn machine. In other words once the machine has chosen a particular action it **increases** the probability of choosing the action irrespective of whether the response from the environment was favourable or unfavourable. However this increase in the action probabilty is done in a systematic and methodical way so that the machine learns the best action which the environment offers in an ε-optimal fashion. The mechanism which we have presented forms an excellent model for an ε-optimal stubbornly learning system.

Apart from the fact that the machine is shown to be ε-optimal, a major contribution of this paper is that the mathematical tools used in this proof (namely the theory of distributions, kernels and topological spaces) are quite distinct from those which are currently used in the field of learning. Besides the above theoretical results, the paper also contains various simulation results which demonstrate the properties of the mechanism presented and which compares it with the traditional L_{RI} scheme

We are currently investigating the properties of multi-action stubbornly learning mechanisms.

Acknowledgements

We would like to thank the Valivetis for preparing the manuscript for us. We are, above all, grateful to Prof. Lakshmivarahan who carefully proofread the paper and critically pre-reviewed it for us.

REFERENCES

[1] Atkinson, R.C., Bower, G.H. and Crothers, E.J., *An Introduction to Mathematical Learning Theory.* New York: Wiley, 1965.

[2] Aso, H., and Kimura, M., "Absolute Expediency of Learning Automata", Information Sciences, Vol. 17, No. 2, 1979, pp.91-112.

[3] Bush, R.R. and Mosteller, F.,*Stochastic Models for Learning.* New York: Wiley, 1958.

[4] Campione, I.C., "The performance of preschool children on reversal and two types of extradimensional shifts," *J. Experimental Child Psychology*, Vol. 11, 1971, pp.480-490.

[5] Dorfman, D.D. and Biderman, M., "A learning model for a continuum of sensory states, " *J. Mathematical Psychology*, Vol. 8, 1971, pp.264-285.

[6] Estes, W.K. and Straughan, J.H., "Analysis of a verbal conditioning situation in terms of statistical learning theory," *J. Experimental Psychology,* Vol. 47, 1954, pp.225-234.

[7] Friedman, M.P., Burke, C.J., Cole, M., Keller, L., Millward, R.B. and Estes, W.K., "Two-choice behavior under extended training with shifting probabilities of reinforcement," *Studies in Mathematical Psychology*, (Ed. by Atkinson, R.C.), Stanford Univ. Press, Stanford, Calif., 1964, pp.250-316.

[8] Fu, K.S., "Learning Control Systems--Review and Outlook", *IEEE Trans. Automatic Control*, Vol. 15, 1970, pp.210-221.

[9] Iosifescu, M. and Theodorescu, R., *Random Processes and Learning.* New York, Springer, 1969.

[10] Krantz, D.H., Atkinson,R.C., Luce, R.D. and Suppes, P., eds., *Contemporary Developments in Mathematical Psychology*, Vol. I, Freeman, San Francisco,1974.

[11] Lakshmivarahan, S., *Learning Algorithms Theory and Applications*, Springer-Verlag, New York, 1981.

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

[13] Lovejoy, E.,"Analysis of the overlearning reversal effect," *Psychological Review,* Vol. 73, 1966, pp.87-103.

[14] Luce, R.D., *Individual Choice Behavior*, New York: Wiley, 1959.

[15] Mendel, J.M. and K.S. Fu, Eds., *Adaptive, Learning and Pattern Recognition Systems*, New York: Academic, 1970.

[16] Mendel, J.M., "Reinforcement learning models and their applications to control problems," *Learning Systems--A Symposium of the 1973 Joint Automatic Control Conf.*, 1973, pp.3-18.

[17] Narendra, K.S., and Thathachar, M.A.L., *Learning Automata*, Prentice-Hall, 1989.

[18] Narendra, K.S., and Thathachar, M.A.L., "Learning Automata -- A Survey", *IEEE Trans. on Syst. Man and Cybernetics*, Vol. SMC-4, 1974, pp.323-334.

[19] Narendra, K.S., and Thathachar, M.A.L., "On the Behaviour of a Learning Automaton in a Changing Environment With Routing Applications", *IEEE Trans.*

on Syst. Man and Cybernetics, Vol. SMC-10, 1980, pp.262-269.

[20] Neveu, J., *Discrete Parameter Matingales,* North-Holland Mathematical Library, Volume10, North-Holland Publishing Company, Amsterdam, 1975.

[21] Norman, M.F., *Markov Processes and Learning Models.* New York: Academic, 1972.

[22] Norman, M.F., "Slow learning", *British J. Math. Statist. Psychology*, Vol. 21, 1968, pp.141-159.

[23] Oommen, B.J., and Thathachar, M.A.L., "Multiaction learning automata possessing ergodicity of the mean", *Information Sciences*, Vol. 35, 1985, pp. 183-198.

[24] Oommen, B.J., "Absorbing and ergodic discretized two-action learning automata", *IEEE Trans. on Syst. Man and Cybernetics*, Vol. SMC-16, 1986, pp.282-296.

[25] Oommen, B. J., and Ma, D. C. Y., "Stochastic automata solutions to the object partitioning problem". To appear in *The Computer Journal.*

[26] Oommen, B. J., "Ergodic Learning Automata Capable of Incorporating *A priori* Information", IEEE Transactions on Systems, Man and Cybernetics., Vol. SMC-17, July/August 1987, pp.717-723.

[27] Paz, A. , *Introduction to Probabilistic Automata.* New York: Academic, 1971.

[28] Ramesh, R., *Learning Automata in Pattern Classification*, M.E. Thesis, Indian Institute of Science, Bangalore, India, 1983.

[29] Sastry, P.S., *Systems of Learning Automata: Estimator Algorithms Applications*, Ph. D. Thesis, Dept. of Electrical Engineering, Indian Institute of Science, Bangalore, India, June 1985.

[30] Thomas, E.A.C, "On a class of additive learning models: Error-correcting and probability matching, " *J. Mathematical Psychology*, Vol. 10, 1973.

[31] Treves, F., Topological Vector Spaces, Distributions and Kernels, Academic Press, Pure and Applied Mathematics Series No.25, New York and London 1967.

[32] Tsetlin, M.L., *Automaton Theory and the Modelling of Biological Systems*, New York and London, Academic, 1973.

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

[34] Tsypkin, Y.Z., *Adaptation and Learning in Automatic Systems.* New York, Academic, 1971.

[35] Varshavskii, V.I., and Vorontsova, I.P., "On the Behaviour of Stochastic Automata With Variable Structure", *Automatica Telemekhanica* (USSR), Vol.24, 1963, pp.327-333.

[36] Viswanathan, R., *Learning Automata : Models and Applications*, Ph. D. Thesis, Dept. of Engineering, Yale University, Conn. , 1972.

[37] Zeaman, D., and House, B.J., "The role of attention in retardate discrimination learning", *Handbook of Mental Deficiency*, N.R. Ellis, ed., McGraw-Hill, New York, 1963, pp.159-223.