Economides, A.A.: **Multiple response learning automata.** *IEEE Transactions on Systems, Man and Cybernetics*, Vol. 26, No 1, pp. 153-156, Febr. 1996.

Multiple Response Learning Automata

Anastasios A. Economides

Abstract— **Learning Automata update their action probabilites on the basis of the response they get from a random environment. They use a reward adaptation rate for a favorable environment's response and a penalty adaptation rate for an unfavorable environment's response. In this correspondence, we introduce Multiple Response learning automata by explicitly classifying the environment responses into a reward (favorable) set and a penalty (unfavorable) set. We derive a new reinforcement scheme which uses different reward or penalty rates for the corresponding reward (favorable) or penalty (unfavorable) responses. Well known learning automata, such as the L_{R-P}, L_{R-I}, $L_{R-\epsilon P}$ are special cases of these Multiple Response learning automata. These automata are feasible at each step, nonabsorbing (when the penalty functions are positive), and strictly distance diminishing. Finally, we provide conditions in order that they are ergodic and expedient.**

I. Introduction

Learning is defined as any relatively permanent change in behavior resulting from past experience, and a learning system is characterized by its ability to improve its behavior with time, in some sense tending toward an ultimate goal [5], [6]. In mathematical psychology, learning systems have been developed to explain behavior patterns among living organizms. These mathematical models in turn have lately been adapted to synthesize engineering systems.

Tsetlin [13] initially introduced the concept of learning automaton operating in an unknown random environment. He considered learning behaviors of finite deterministic automata under a stationary random environment. Varshavskii & Vorontsova [14] introduced variable structure stochastic automata in an unknown random environment. A number of papers have appeared recently [1], [3], [4], [7], [8]–[12] that propose new reinforcement schemes for learning automata and investigate their properties. Thathachar and Sastry [12] introduce estimates of the environment reward characteristics into the learning automata algorithms. Oommen and Thathachar [10] prove necessary and sufficient conditions for ergodicity in the mean of learning automata. Oommen [7], [9] introduces discretized learning automata, where the action probability assumes one of a finite number of distinct values in [0, 1]. He also [8] presents ergodic learning automata which can take into consideration a priori information about the action probabilities. Simha and Kurose [11] propose learning automata based on the relative reward strength of the environment response. Finally, we [1], [3], [4] introduce several learning automata algorithms for adaptively routing newly arriving calls in virtual circuit networks, as well in packet switched networks with error prone links.

Previous studies on P-model learning automata algorithms assume only two possible updating schemes. If the environment response is favorable (success), then we reward the selected action, while if the environment response is unfavorable (failure), then we penalize the selected action.

In this correspondence, we extend the general L_{R-P} scheme to Q-models by introducing different updating parameters for different environment responses for the selected action. We reward (or penalize) the selected action according to how much favorable (or unfavorable) the environment response was. If the environment response is very favorable, the selected action is rewarded very much. On the other hand, if the environment response is marginally favorable, the selected action is penalized very little. Similarly, if the environment response is very unfavorable, the selected action is heavily penalized, and if the environment response is marginally unfavorable, the selected action is marginally penalized. These concepts lead to a generalization of the traditional learning automata algorithms.

Manuscript received July 1, 1993; revised April 2, 1994, September 1, 1994, and February 20, 1995.

The author is with the University of Macedonia, Thessaloniki 54006, Greece.

Publisher Item Identifier S 1083-4419(96)00423-2.

II. Multiple Response Learning Automata

A learning automaton is a feedback system connecting a stochastic automaton and an environment. At each instant n, the automaton selects an action $a(n) = a_i$ (among actions $a_1, \ldots, a_{|a|}$) with probability $P_i(n) = P[a(n) = a_i]$. Let $\mathbf{P}(n) = [P_1(n), \ldots, P_{|a|}(n)]$ be the vector of action probabilities. Action $a(n)$ becomes input to the environment. If this results in a favorable environment response $(X(n) \rightarrow 0)$, then the probability $P_i(n)$ is increased by $\Delta P_i(n)$ and the $P_j(n)$, $j \neq i$, are decreased by $\Delta P_j(n)$. Otherwise, if an unfavorable environment response $(X(n) \rightarrow 1)$ appears, then the $P_i(n)$ is decreased by $\bar{\Delta} P_i(n)$ and the $P_j(n)$, $j \neq i$ are increased by $\bar{\Delta} P_j(n)$. By iteration of the algorithm, we achieve adaptation to varying environment conditions.

In this correspondence, we introduce Multiple Response (MR) learning automata algorithms. The idea is to use different adaptation rates for different environment responses. Whenever the environment response is very good $(X(n) \rightarrow 0)$ (reward response 1), we heavily reward the selected action by increasing its probability rapidly. When the environment response is marginally good (reward response R), we correspondingly reward the selected action by increasing its probability slowly. Analogously, whenever the environment response is very bad $(X(n) \rightarrow 1)$ (penalty response 1), then we heavily penalize the selected action by decreasing its probability very fast. When the cost of the environment response is not so bad (penalty response P), we penalize the selected action less strictly by decreasing its probability slowly.

Next, we introduce a Q-model Multiple Response (Q-MR) learning automaton algorithm, for which the environment response takes discrete values, normalized to be in the [0, 1] interval. So, if action a_i was selected at time n, the environment response is an element of the set $\{X_i^1, \ldots, X_i^R, \bar{X}_i^P, \ldots, \bar{X}_i^1\}$, i.e.

Let $a(n) = a_i$

reward response 1:	$X(n) = X_i^1$
reward response 2:	$X(n) = X_i^2$
...	
reward response R:	$X(n) = X_i^R$
penalty response P:	$X(n) = \bar{X}_i^P$
penalty response $P-1$:	$X(n) = \bar{X}_i^{P-1}$
...	
penalty response 1:	$X(n) = \bar{X}_i^1$

where

$$0 \leq X_i^1 < X_i^2 < \cdots < X_i^R < m_i$$
$$< \bar{X}_i^P < \bar{X}_i^{P-1} < \cdots < \bar{X}_i^1 \leq 1.$$

A possible sequence for the reward responses $\{X_i^r\}$ could be a Fibonacci sequence (normalized to the $[0, m_i)$ interval). Also a possible sequence for the penalty responses $\{\bar{X}_i^p\}$ could be a Fibonacci sequence (normalized to the $(m_i, 1]$ interval). The m_i is the threshold for a response to be considered as reward or penalty.

Responses below this threshold are considered as rewards, while responses above it are considered as penalties.

If the selected action a_i results in good environment response $(0 \leq X(n) < m_i)$, then we reward this action; Otherwise $(m_i < X(n) \leq 1)$, we penalize it. However, as opposed to traditional schemes, the reward (penalty) parameters depend on how good (bad) the environment response is. Therefore, for each of the above environment responses, we use different reward functions $g_i^r(.)$, $r = 1, \ldots, R$ and penalty functions $h_i^p(\cdot)$, $p = 1, \ldots, P$, with $1 > g_i^1(\cdot) > g_i^2(\cdot) > \cdots > g_i^R(\cdot) > 0$, and $1 > h_i^1(\cdot) > h_i^2(\cdot) > \cdots > h_i^P(\cdot) > 0$. These are functions of the environment state measurements $\mathbf{X}(n)$.

The above concepts lead us to formally defining the *Q-MR algorithm*:

$$
\begin{aligned}
&\text{Let } a(n) = a_i\\
&\text{For } r = 1 \text{ to } R\\
&\quad \text{If } X(n) = X_i^r, \text{ then}\\
&\quad\quad P_i(n+1) = P_i(n) + g_i^r(\mathbf{X}(n))[1 - P_i(n)]\\
&\quad\quad P_j(n+1) = P_j(n) - g_i^r(\mathbf{X}(n))P_j(n) \quad \forall j \neq i
\end{aligned}
$$

$$
\begin{aligned}
&\text{For } p = 1 \text{ to } P\\
&\quad \text{If } X(n) = \bar{X}_i^p, \text{ then}\\
&\quad\quad P_i(n+1) = P_i(n) - h_i^p(\mathbf{X}(n))P_i(n)\\
&\quad\quad P_j(n+1) = P_j(n) + h_i^p(\mathbf{X}(n))\\
&\quad\quad\quad \times \left[\frac{1}{|a|-1} - P_j(n)\right] \quad \forall j \neq i
\end{aligned}
$$

For the special case of $g_i^r(\cdot) = \theta * \alpha_i^r$ and $h_i^p(\cdot) = \theta * \beta_i^p$, with $0 < \theta \leq 1, 0 < \alpha_i^r < 1, 0 < \beta_i^p < 1$, we have the *Q-model Multiple Response Linear (Q-MRL) algorithm*:

$$
\begin{aligned}
&\text{Let} a(n) = a_i\\
&\text{For} r = 1 \text{ to } R\\
&\text{If} X(n) = X_i^r, \text{ then}\\
&\quad P_i(n+1) = P_i(n) + \theta\alpha_i^r[1 - P_i(n)]\\
&\quad P_j(n+1) = P_j(n) - \theta\alpha_i^r P_j(n) \quad \forall j \neq i
\end{aligned}
$$

$$
\begin{aligned}
&\text{For } p = 1 \text{ to } P\\
&\text{If } X(n) = \bar{X}_i^p, \text{ then}\\
&\quad P_i(n+1) = P_i(n) - \theta\beta_i^p P_i(n)\\
&\quad P_j(n+1) = P_j(n) + \theta\beta_i^p\\
&\quad\quad \times \left[\frac{1}{|a|-1} - P_j(n)\right] \quad \forall j \neq i
\end{aligned}
$$

where the reward rates α_i^r, $r = 1, \ldots, R$ and penalty rates β_i^p, $p = 1, \ldots, P$ satisfy $1 > \alpha_i^1 > \alpha_i^2 > \cdots > \alpha_i^R > 0$, and $1 > \beta_i^1 > \beta_i^2 > \cdots > \beta_i^P > 0$.

Define also the *Q-MRL$_{\alpha=\beta}$ algorithm*, when $R = P$ and $\alpha_i^k = \beta_i^k$ $k = 1, \ldots, R$, the *Q-MRL$_{\beta=\epsilon\alpha}$ algorithm*, when $R = P$ and $\beta_i^k = \epsilon_i^k \alpha_i^k, < \epsilon_i^k \ll 1$ $k = 1, \ldots, R$, and the *Q-MRL$_\alpha$ algorithm*, when $\beta_i^p = 0$ $p = 1, \ldots, P$. Note that for the special cases of $R = P = 1$, we have the L_{R-P}, $L_{R-\epsilon P}$ and L_{R-I} algorithms.

In this section, we introduce the Multiple Response (MR) learning automata algorithms. In [1] and [2], we have simulated MR learning automata algorithms for routing in computer networks. Next, we will investigate the behavior of these algorithms.

III. Norms of Behavior

In this section, we provide a quantitative basis for evaluating the performance of MR learning automata. We state the definitions of some useful norms of behavior, such as expediency, optimality, and ϵ-optimality. The MR algorithms are feasible, and stricly distance diminishing. Finally, we state conditions in order to be ergodic and expedient.

Let $d_i^r = P[X(n) = X_i^r / a(n) = a_i] \in (0,1)$ be the unknown probability for reward response r, when action a_i is selected, and $c_i^p = P[X(n) = \bar{X}_i^p / a(n) = a_i] \in (0,1)$ be the unknown probability for penalty response p, when action a_i is selected, such that

$$\sum_{r=1}^{R} d_i^r + \sum_{p=1}^{P} c_i^p = 1.$$

If at a certain time instant n, the automaton selects action $a(n)$ with probability $\mathbf{P}(n)$, then the *average response* received by the automaton conditioned on $\mathbf{P}(n)$ is

$$
\begin{aligned}
M(n) &= E[X(n)/\mathbf{P}(n)]\\
&= \sum_{i=1}^{|a|} E[X(n)/\mathbf{P}(n), a(n) = a_i] P_i(n)\\
&= \sum_{i=1}^{|a|} \left[\sum_{r=1}^{R} X_i^r d_i^r + \sum_{p=1}^{P} \bar{X}_i^p c_i^p\right] P_i(n) \qquad (1)
\end{aligned}
$$

the *average reward* received by the automaton conditioned on $\mathbf{P}(n)$ is

$$M^R(n) = \sum_{i=1}^{|a|} \sum_{r=1}^{R} X_i^r d_i^r P_i(n) \qquad (2)$$

and the *average penalty* received by the automaton conditioned on $\mathbf{P}(n)$ is

$$M^P(n) = \sum_{i=1}^{|a|} \sum_{p=1}^{P} \bar{X}_i^p c_i^p P_i(n) \qquad (3)$$

Thus, we can write $M(n) = M^R(n) + M^P(n)$.

Taking expectations and then the limit as $n \to \infty$, we have

$$\lim_{n\to\infty} E[M(n)] = \sum_{i=1}^{|a|} \left[\sum_{r=1}^{R} X_i^r d_i^r + \sum_{p=1}^{P} \bar{X}_i^p c_i^p\right] \lim_{n\to\infty} E[P_i(n)]. \qquad (4)$$

If no a priori information is available and the actions are chosen at random $(P_1(n) = \cdots = P_{|a|}(n) = 1/|a|)$, then the average response received by the automaton conditioned on $\mathbf{P}(n)$ is

$$M_0 = \frac{1}{|a|} \sum_{i=1}^{|a|} \left[\sum_{r=1}^{R} X_i^r d_i^r + \sum_{p=1}^{P} \bar{X}_i^p c_i^p\right]. \qquad (5)$$

We shall use this *pure-chance automaton* as the standard for comparison. An automaton that performs better than the pure-chance automaton is said to be expedient.

Definition 1: A MR learning automaton is called *expedient* iff $\lim_{n\to\infty} E[M(n)] < M_0$.

However, we are interested in automata that exhibit much better behavior. It would be desirable if $\lim_{n\to\infty} E[M(n)]$ could be minimized by a proper selection of the actions. We say that an automaton whose average response tends to its minimum value is optimal.

Definition 2: A MR learning automaton is called optimal iff

$$\lim_{n\to\infty} E[M(n)] = \min_i \left\{ \sum_{r=1}^{R} X_i^r d_i^r + \sum_{p=1}^{P} \bar{X}_i^p c_i^p \right\}.$$

The optimality implies that the action a_i associated with the minimum $\sum_{r=1}^{R} X_i^r d_i^r + \sum_{p=1}^{P} \bar{X}_i^p c_i^p$ (average response to action a_i) is chosen asymptotically with probability one. However, there are environments where it is impossible to achieve optimality. In such cases, a suboptimal solution may be acceptable. ϵ-optimality represents one such suboptimal behavior.

Definition 3: A MR learning automaton is called ϵ-optimal iff

$$\lim_{n\to\infty} E[M(n)] < \min_i \left\{ \sum_{r=1}^{R} X_i^r d_i^r + \sum_{p=1}^{P} \bar{X}_i^p c_i^p \right\} + \epsilon \qquad \epsilon > 0.$$

$\epsilon-$optimality implies that the performance of the automaton can be made as close to the optimal as desired.

The Multiple Response learning automata algorithms have two useful properties. First of all, the Q-MR algorithm preserves the *feasibility* of the action probability space, i.e. at each iteration of the MR algorithm, the action probabilities are always nonnegative and sum to 1. The second result is that the Q-MR algorithm (with positive penalty functions) is *nonabsorbing*, i.e. it is not trapped in a specific action (no action is selected with probability 1). This is a desirable property for dynamic systems where the optimal action continuously changes over time. In this case, an action that was optimal at a given moment may not be optimal any more, and so we like to give a chance to the other actions.

Furthermore, at each step, the Q-MR algorithm approach to the optimal action. In other words, successive applications of the algorithm on two different trajectories of the action probabilities bring them closer to each other and finally result in convergence to the asymptotic action probabilities.

Theorem 1: The Q-MR algorithm is strictly distance diminishing.

The proof of the Theorem 1 is similar to corresponding material in [5]. The above properties lead to the ergodic character of the Q-MR algorithm. Hence, the sequence $\{\mathbf{P}(n)\}$ converges in distribution to a random variable $\mathbf{P}^*$. The following theorem states this more formally.

Theorem 2: The Q-MRL algorithm with $\sum_{r=1}^{R} \alpha_i^r d_i^r + \sum_{p=1}^{P} \beta_i^p c_i^p =$ constant $\forall i$ is ergodic and $\mathbf{P}(n)$ converges in distribution to a random variable with mean

$$\lim_{n\to\infty} E[P_i(n)] = \frac{\frac{1}{\sum_{p=1}^{P} \beta_i^p c_i^p}}{\sum_{j=1}^{|a|} \frac{1}{\sum_{p=1}^{P} \beta_j^p c_j^p}} \forall i$$

independent of the initial probability $\mathbf{P}(0)$.

Proof: It is required to prove that the following equation holds

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

where $\mathbf{Q}$ is a stochastic matrix with no absorbing barriers. The conditional expectation of $P_i(n+1)$ given $\mathbf{P}(n)$ is

$$\begin{aligned}
&E[P_i(n+1)/\mathbf{P}(n)] \\
&= \sum_{r=1}^{R} [P_i(n) + \theta\alpha_i^r[1 - P_i(n)]]P_i(n)d_i^r \\
&\quad + \sum_{p=1}^{P} [P_i(n) - \theta\beta_i^p P_i(n)]P_i(n)c_i^p \\
&\quad + \sum_{j=1 j\neq i}^{|a|} \sum_{r=1}^{R} [P_i(n) - \theta\alpha_j^r P_i(n)]P_j(n)d_j^r \\
&\quad + \sum_{j=1 j\neq i}^{|a|} \sum_{p=1}^{P} \left[P_i(n) + \theta\beta_j^p \left[\frac{1}{|a|-1} - P_i(n)\right]\right] P_j(n)c_j^p \\
&= P_i(n) + \theta P_i(n) \sum_{r=1}^{R} \left[\alpha_i^r[1 - P_i(n)]d_i^r - \sum_{j=1 j\neq i}^{|a|} \alpha_j^r P_j(n)d_j^r\right] \\
&\quad + \theta \sum_{p=1}^{P} \left[\sum_{j=1 j\neq i}^{|a|} \beta_j^p \left[\frac{1}{|a|-1} - P_i(n)\right] P_j(n)c_j^p - \beta_i^p [P_i(n)]^2 c_i^p\right] \\
&= P_i(n) + \theta P_i(n) \sum_{r=1}^{R} \left[\alpha_i^r \sum_{j=1 j\neq i}^{|a|} P_j(n)d_i^r - \sum_{j=1 j\neq i}^{|a|} \alpha_j^r P_j(n)d_j^r\right] \\
&\quad + \theta \sum_{p=1}^{P} \left[\sum_{j=1 j\neq i}^{|a|} \beta_j^p \left[\frac{1}{|a|-1} - P_i(n)\right] P_j(n)c_j^p - \beta_i^p [P_i(n)]^2 c_i^p\right] \\
&= P_i(n) + \theta P_i(n) \sum_{r=1}^{R} \sum_{j=1 j\neq i}^{|a|} P_j(n)(\alpha_i^r d_i^r - \alpha_j^r d_j^r) \\
&\quad + \theta \sum_{p=1}^{P} \left[\sum_{j=1 j\neq i}^{|a|} \beta_j^p \left[\frac{1}{|a|-1} - P_i(n)\right] P_j(n)c_j^p - \beta_i^p [P_i(n)]^2 c_i^p\right].
\end{aligned}$$

It is well known that even the special cases of the present algorithms (the traditional L_{R-P}) has quadratic terms whenever $\alpha \neq \beta$. One cannot, of cource, expect it to be any simpler in this case. The expressions obtained are cumbersome and specialize only for the special case where the assumption of the Theorem holds.

Taking expectations on both sides, we have

$$\begin{aligned}
&E[P_i(n+1)] \\
&= E[P_i(n)] + \theta \sum_{r=1}^{R} \sum_{j=1,j\neq i}^{|a|} E[P_i(n)P_j(n)](\alpha_i^r d_i^r - \alpha_j^r d_j^r) \\
&\quad + \frac{\theta}{|a|-1} \sum_{p=1}^{P} \sum_{j=1,j\neq i}^{|a|} E[P_j(n)]\beta_j^p c_j^p \\
&\quad - \theta \sum_{p=1}^{P} \sum_{j=1,j\neq i}^{|a|} E[P_i(n)P_j(n)]\beta_j^p c_j^p \\
&\quad - \theta \left[E[P_i(n)] - \sum_{j=1,j\neq i}^{|a|} E[P_i(n)P_j(n)]\right] \sum_{p=1}^{P} \beta_i^p c_i^p \\
&= E[P_i(n)] + \theta \sum_{j=1,j\neq i}^{|a|} E[P_i(n)P_j(n)] \sum_{r=1}^{R} (\alpha_i^r d_i^r - \alpha_j^r d_j^r) \\
&\quad + \frac{\theta}{|a|-1} \sum_{j=1,j\neq i}^{|a|} E[P_j(n)] \sum_{p=1}^{P} \beta_j^p c_j^p \\
&\quad - \theta \sum_{j=1,j\neq i}^{|a|} E[P_i(n)P_j(n)] \sum_{p=1}^{P} \beta_j^p c_j^p \\
&\quad - \theta E[P_i(n)] \sum_{p=1}^{P} \beta_i^p c_i^p + \theta \sum_{j=1,j\neq i}^{|a|} E[P_i(n)P_j(n)] \sum_{p=1}^{P} \beta_i^p c_i^p
\end{aligned}$$

$$\sum_{i=1}^{|a|} \frac{\sum_{j=1}^{|a|} \left[\left(\sum_{r=1}^{R} X_i^r d_i^r + \sum_{p=1}^{P} \bar{X}_i^p c_i^p\right) - \left(\sum_{r=1}^{R} X_j^r d_j^r + \sum_{p=1}^{P} \bar{X}_j^p c_j^p\right)\right]}{\sum_{p=1}^{P} \beta_i^p c_i^p} < 0$$

$$= \left[1 - \theta \sum_{p=1}^{P} \beta_i^p c_i^p\right] E[P_i(n)]$$
$$+ \theta \sum_{j=1, j\neq i}^{|a|} E[P_i(n)P_j(n)] \left[\sum_{r=1}^{R} (\alpha_i^r d_i^r - \alpha_j^r d_j^r) + \sum_{p=1}^{P} (\beta_i^p c_i^p - \beta_j^p c_j^p)\right]$$
$$+ \frac{\theta}{|a|-1} \sum_{j=1, j\neq i}^{|a|} E[P_j(n)] \sum_{p=1}^{P} \beta_j^p c_j^p$$

Observe that in general, $E[P_i(n+1)]$ is not linear in $\mathbf{P}$. However, with the assumption

$$\sum_{r=1}^{R} \alpha_i^r d_i^r + \sum_{p=1}^{P} \beta_i^p c_i^p = \text{ constant } \forall i \Rightarrow$$
$$\left[\sum_{r=1}^{R} (\alpha_i^r d_i^r - \alpha_j^r d_j^r) + \sum_{p=1}^{P} (\beta_i^p c_i^p - \beta_j^p c_j^p)\right] = 0 \quad \forall i, j$$

the quadratic terms disappear, yielding $E[\mathbf{P}(n+1)] = \mathbf{Q}^T E[\mathbf{P}(n)]$, where

$$q_{ii} = 1 - \theta \sum_{p=1}^{P} \beta_i^p c_i^p$$
$$q_{ji} = \frac{\theta \sum_{p=1}^{P} \beta_j^p c_j^p}{|a|-1}$$

with $q_{ii}, q_{ij} \in [0, 1)\ \forall i, j = 1, \ldots, |a|$ and $\sum_{j=1}^{|a|} q_{ij} = 1\ \forall i$

$\mathbf{Q}$ is a stochastic matrix with no absorbing barrier, hence ergodic. Thus, the limiting value of $E[\mathbf{P}(n)]$ exists, is independent of $\mathbf{P}(0)$ and is given as the solution of the equation $\mathbf{P}^* = \mathbf{Q}^T\mathbf{P}^*$.

Then the limiting probabilities are

$$\lim_{n\to\infty} E[P_i(n)] = \frac{\frac{1}{\sum_{p=1}^{P} \beta_i^p c_i^p}}{\sum_{j=1}^{|a|} \frac{1}{\sum_{p=1}^{P} \beta_j^p c_j^p}}$$

□

Thus, if $\sum_{p=1}^{P} \beta_i^p c_i^p < \sum_{p=1}^{P} \beta_j^p c_j^p\ \forall j \neq i$, then $\lim_{n\to\infty} E[P_i(n)] > \lim_{n\to\infty} E[P_j(n)]\ \forall j \neq i$, i.e. if the sum of the penalty probability rates for action a_i is smaller than that of any other action a_j, then on the average, action a_i is chosen asymptotically with a higher probability than any other action a_j.

The ergodicity condition

$$\sum_{r=1}^{R} \alpha_i^r d_i^r + \sum_{p=1}^{P} \beta_i^p c_i^p = \text{ constant } \forall i$$

says that, the sum of the reward probability rates $(\alpha_i^r d_i^r)$ plus the sum of the penalty probability rates $(\beta_i^p c_i^p)$ should be the same for all actions.

Thus, if $\sum_{r=1}^{R} \alpha_i^r d_i^r > \sum_{r=1}^{R} \alpha_j^r d_j^r\ \forall j \neq i$, then $\lim_{n\to\infty} E[P_i(n)] > \lim_{n\to\infty} E[P_j(n)]\ \forall j \neq i$, i.e. if the sum of the reward probability rates for action a_i is larger than that of any other action a_j, then on the average, action a_i is chosen asymptotically with a higher probability than any other action a_j.

For the simple case of two actions with $R = P = 1$, $\alpha_i = \alpha$ and $\beta_i = \beta\ \forall i$, the ergodicity condition becomes $\alpha = \beta$, i.e. the reward rate equals the penalty rate. This is the case of the traditional L_{R-P} learning automaton.

Finally, the Q-MRL algorithm is expedient under more general conditions than the L_{R-P}-algorithm.

Theorem 3: The Q-MRL algorithm with the conditions of Theorem 2 and (see the equation at the bottom of the previous page) is expedient.

Proof: Using the asymptotic action probabilities from Theorem 2 and the condition of Theorem 3, it is easy to show that $\lim_{n\to\infty} E[M(n)] < M_0$. □

IV. Conclusion

In this correspondence, we suggested a generalization of the traditional learning automata algorithms. We introduce Multiple Response (MR) learning automata algorithms that use different adaptation rates for different environment responses. The well known $L_{R-I}, L_{R-P}, L_{R-\epsilon P}$ learning automata are special cases of this new class of learning automata. These MR algorithms are feasible at each step, nonabsorbing (when the penalty functions are positive) and strictly distance diminishing. Finally, we provide conditions in order that they are ergodic and expedient. For applications of these MR algorithms to problems in distributed systems see [1] and [2]. An open problem for future research is the proper selection of the reward and penalty rates to satisfy the ergodicity condition.

Acknowledgment

The author is grateful to the anonymous reviewers for comments which gave me a greater insight into the field.

References

[1] A. A. Economides, "Learning automata routing in connection-oriented networks," *Int. J. Commun. Syst.*, vol. 8, pp. 225–237, 1995.

[2] ——, "A unified game-theoretic methodology for the joint load sharing, routing and congestion control problem," *Ph.D. Dissertation*, University of Southern California, Los Angeles, Aug. 1990.

[3] A. A. Economides, P. A. Ioannou and J. A. Silvester, "Decentralized adaptive routing for virtual circuit networks using stochastic learning automata," in *Proc. IEEE Infocom 88 Conf.*, 1988, pp. 613–622.

[4] A. A. Economides and J. A. Silvester, "Optimal routing in a network with unreliable links," in *Proc. IEEE Computer Networking Symp.*, 1988, pp. 288–297.

[5] K. Narendra and M. A. L. Thathacher, *Learning Automata: An Introduction*. Englewood Cliffs, NJ: Prentice Hall, 1989.

[6] ——, "Learning automata: A survey," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-4, no. 4, pp. 323–334, July 1974.

[7] B. J. Oommen, "Absorbing and ergodic discretized two-action learning automata," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-16, no. 2, pp. 282–293, Mar./Apr. 1986.

[8] ——, "Ergodic learning automata capable of incorporating a priori information," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-17, no. 4, pp. 717–723, July/Aug. 1987.

[9] B. J. Oommen and J. P. R. Christensen, "ϵ-optimal discretized linear reward-penalty learning automata," *IEEE Trans. Syst., Man, Cybern.*, vol. 18, no. 3, pp. 451–458, May/June 1988.

[10] B. J. Oommen and M. A. L. Thathachar, "Multiaction learning automata possessing ergodicity of the mean," *Inform. Sci.*, vol. 35, pp. 183–198, 1985.

[11] R. Simha and J. F. Kurose, "Relative reward strength algorithms for learning automata," *IEEE Trans. Syst., Man, Cybern.*, vol. 19, no. 2, pp. 388–398, Mar./Apr. 1989.

[12] M. A. L. Thathachar and P. S. Sastry, "A new approach to the design of reinforcement schemes for learning automata," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-15, no. 1, pp. 168–175, Jan./Feb. 1985.

[13] M. L. Tsetlin, "On the behavior of finite automata in random media," *Avtomatika i Telemekhanika*, vol. 22, no. 10, pp. 1345–1354, Oct. 1961.

[14] V. I. Varshavskii and I. P. Vorontsova, "On the behavior of stochastic automata with a variable structure," *Avtomatika i Telemekhanika*, vol. 24, no. 3, pp. 353–360, Mar. 1963.