INFORMATION SCIENCES **28**, 1–20 (1982)

ε-Optimality of a General Class of Learning Algorithms*

M. R. MEYBODI
and
S. LAKSHMIVARAHAN

School of Electrical Engineering and Computer Science
University of Oklahoma, Norman, Oklahoma 73019

Communicated by John M. Richardson

ABSTRACT

Conditions of ε-optimality of a general class of strongly absolutely expedient learning algorithms are derived.

1. INTRODUCTION

Learning algorithms have been extensively studied in mathematical psychology [13, 14, 18] for decades, and more recently in learning automata theory and mathematical statistics [1, 2, 8, 11, 16]. In mathematical psychology the interest in learning stems from the desire to understand the observed learning in animals and associated changes in their behavior. However, in learning automata theory and mathematical statistics the aim is to build algorithms that exhibit prespecified behaviour. Our interests, in this paper, are in the latter approach.

We study a new class of absorbing barrier algorithms of the reward-penalty type which display identical behavior under the occurrence of success and failure. Necessary and sufficient conditions for strong absolute expediency of this class of algorithms are given. As a consequence, ε-optimality of this class of algorithm is obtained.

*Part of the results in this paper were presented in the conference on "Mathematical Learning Models—Theory and Applications" held at the University of Bonn, West Germany during 3–7 May 1982.

52 Vanderbilt Ave., New York, NY 10017 0020-0255/82/07001-20$02.75

The concept of absolute expediency was originally introduced by Lakshmivarahan and Thathachar [3] in 1973. Since then this concept has played a major role in the analysis and design of ε-optimal absorbing barrier algorithms, [1, 2, 4, 5, 7]. The results of this paper shed further light on this concept. Note that ε-optimal nonabsorbing learning algorithms have also been recently developed in [15] and discussed extensively in a recent monograph by Lakshmivarahan [21].

A number of learning algorithms of the reward-penalty type whose behavior is asymmetrical with respect to the occurrence of success and failure have been extensively studied by many researchers—Varshavskii and Vorantsova [10], Fu [8], McMurtry and Fu [9], Shapiro and Narendra [20], Viswanathan and Narendra [19], Chandrasekaran and Shen [17], Lakshmivarahan and Thathachar [3], and Sawaragi and Baba [7], to mention a few. For a class of learning algorithms which are similar to the one discussed here, Aso and Kimura [5] derived necessary and sufficient conditions for absolute expediency [3]. Interestingly, Aso and Kimura [5] call this class of algorithms "stochastic vector automation" algorithms. Recently, for the same class of algorithms considered in this paper, Herkenrath, Kalin, and Lakshmivarahan [6] derived necessary and sufficient conditions for absorption at the vertices of the unit simplex of proper dimension.

In Section 2, the algorithm and the statement of problem are given. Necessary and sufficient conditions for strong absolute expediency and convergence of the algorithm with probability one are established in Section 3. ε-optimality is derived in Section 4. A number of computer simulation results are given in Section 5.

2. LEARNING ALGORITHM AND STATEMENT OF PROBLEM

There are M $(2 \leqslant M < \infty)$ coins. At time instant k $(= 0,1,2,\ldots)$, coin i $(1 \leqslant i \leqslant M)$ is chosen for tossing with probability $P_i(k)$, where

$$P(k) = (P_1(k), P_2(k), \ldots, P_M(k))^T,$$

$\sum_{i=1}^{M} P_i(k) = 1$, $0 \leqslant P_i(k)$, and T denotes transpose. In tossing, the ith coin falls heads (tails) with probability d_i $(c_i = 1 - d_i)$. It is *assumed* that (1) $0 < d_i < 1$, $i = 1,2,\ldots,M$; (2) the d_i's are all distinct (that is, $d_i \neq d_j$ for all $i \neq j$); the (3) d_i's do not depend on k; and (4) the d_i's are all *unknown*. The outcome of falling heads is called *success*, and falling tails is called *failure*. Let $D = (d_1, d_2, \ldots, d_M)^T$, and without loss of generality assume

$$d_1 > d_2 > d_3 > \cdots > d_M. \tag{1}$$

Any vector D that satisfies the above conditions is called an *admissible* D. The average probability of success at state k, denoted by $\eta(k)$, is

$$\eta(k) = \sum_{i=1}^{M} P_i(k)\, d_i. \tag{2}$$

Our basic problem is to make the average probability of success as close to its maximum (that is, d_1) as possible. The primary interest in and the challenge of this problem arise basically from the fact that the success probabilities (d_i's) of the coin are unknown.

As a first step towards the solution to this problem, in this paper, we propose to change $p(k)$ in a learning algorithm. The key idea of the learning algorithm may be stated in words as follows: Increase (decrease) the probability of choosing the ith coin at time $k+1$, if it was chosen for tossing at time k and the toss resulted in success (failure). In particular, let

$$S_M = \left\{ P | P = (P_1, P_2, \ldots P_M)^T, 0 \leqslant P_i, \sum_{i=1}^{M} P_i = 1 \right\}$$

be the M-dimensional unit simplex, and let

$$V_M = \{ e_i | i = 1,2,\ldots,M \}, \qquad \text{where} \quad e_i = (0,0,0,\ldots 1,\ldots 0)^T,$$

be the ith unit vector of dimension M. Further, let

$$S_M^0 = \left\{ P | P = (P_1, P_2, \ldots, P_M)^T, 0 < P_i, \sum_{i=1}^{M} P_i = 1 \right\}$$

Clearly, V_M corresponds to the corners of vertices of S_M. S_M^0 is called the interior of S_M, and $\delta S_M = S_M - S_M^0$ is called the boundary of S_M (where $A - B$ is called the relative complement of the set B with respect to set A). Let $f_s^i[\ ], g_s^i[\ ]$ be continuous functions such that

$$f_s^i : S_M \to [0,1], \quad g_s^i : S_M \to [0,1], \qquad i, s = 1,2,\ldots,M.$$

Formally the above algorithms may be stated as follows:

$$\begin{aligned} P_s(k+1) &= P_s(k) - \theta f_s^i[P(k)], \qquad s \neq i, \\ P_i(k+1) &= P_i(k) + \theta f_i^i[P(k)], \end{aligned} \tag{3a}$$

if the toss of coin i resulted in success, and

$$P_i(k+1) = P_i(k) - \theta g_i^i[P(k)],$$
$$P_s(k+1) = P_s(k) + \theta g_s^i[P(k)], \qquad s \neq i, \tag{3b}$$

if the toss of coin i resulted in failure, where $0 < \theta \leqslant 1$ is called the *step length parameter* and the following consistency condition is imposed: if

either $$f_s^i[p] \equiv 0 \qquad \text{for all } p \in S_M$$

or $$f_s^i[p] \leqslant P_s, \quad s \neq i, \quad \text{and} \quad f_i^i[p] = \sum_{s \neq i} f_s^i[p] \tag{4a}$$

and

either $$g_s^i[p] \equiv 0 \qquad \text{for all} \quad p \in S_M$$

or $$g_i^i[p] \leqslant p_i, \quad \text{and} \quad g_i^i[p] = \sum_{s \neq i} g_s^i[p] \tag{4b}$$

for all $i, s = 1,2,\ldots M$, then $p(k+1) \in S_M$ if $p(k)$ does. However, in order to make the algorithm nontrivial and interesting, either $f_s^i \equiv 0$ or $g_s^i \equiv 0$ but not both.

REMARK 1. If coin i is chosen for tossing, it is clear from (3) that success (failure) increases (decreases) the probability of its choice. The increase and decrease are called reward and penalty, and hence (3) is called a reward-penalty algorithm. If $f_s^i \not\equiv 0$ but $g_s^i \equiv 0$, then (3) is called a reward-inaction algorithm; if $f_s^i \equiv 0$ but $g_s^i \not\equiv 0$, then it is called an inaction-penalty algorithm.

Since the vector D of success probabilities and the functions $f_s^i[\ ]$ and $g_s^i[\ ]$ (for all i and s) are independent of k, the sequence $\{p(k)\}$, $k \geqslant 0$, is clearly a discrete time Markov process with stationary transition function over the state space S_M. To quantify the behavior of the process $\{p(k)\}$, and hence of $\{\eta(k)\}$, we introduce some definitions. Let

$$I = \{1,2,\ldots,M\}, \qquad E = \{\text{success, failure}\}.$$

The algorithm (3) defines a mapping $T: S_M \times I \times E \to S_M$, where

$$P(k+1) = T[p(k), i(k), e(k)], \tag{5}$$

$i(k) \in I$ is the coin chosen for tossing, and $e(k) \in E$ is the outcome of the toss of that coin at time k.

DEFINITION 1. A state $p \in S_M$ is said to be *absorbing state* if and only if

$$p = T[p, i, e] \qquad \text{with probability one.} \tag{6}$$

DEFINITION 2. A learning algorithm T is said to be *absorbing* if and only if there is at least one absorbing state.

For reasons that will become apparent, in this paper our interests are in the class of learning algorithms with V_M as the only set of absorbing states.

DEFINITION 3. A learning algorithm T is said to be

(a) *optimal* if

$$\lim_{k \to \infty} E[\eta(k)] = d_1, \tag{7}$$

(b) ε-*optimal* if for all $p(0) = p \in S_M^0$

$$\lim_{k \to \infty} |E[\eta(k)] - d_1| < \varepsilon, \qquad \varepsilon > 0, \tag{8}$$

(c) *absolutely expedient* if

$$E[\eta(k+1) | p(k) = p] \geqslant \eta(k) \qquad \text{with probability one} \tag{9}$$

for all admissible D (satisfying the assumptions given at the beginning of this section) and for all $p \in S_M$, with strict inequality in (9) holding good for all $p \in S_M^0$,

(d) *strongly absolutely expedient* if

$$E[\eta(k+1) | p(k) = p] \geqslant \eta(k) \qquad \text{with probability one} \tag{10}$$

for all admissible D and for all $p \in S_M$, with strict inquality in (10) holding good for all $p \in (S_M - V_M)$

STATEMENT OF PROBLEM. Our aim in this paper is to find conditions for ε-optimality of the general class of learning algorithm given in (3).

REMARK 2. Algorithm (3) is a generalization of many known algorithms in the sense that the functions in (3) used for updating the probabilities depend on the coin chosen as well. Such generalized algorithms are considered in [5] and [6]. In most of the earlier papers the functions that are used in updating are independent of the coin that is chosen for tossing [1–4, 7–9, 10, 12, 13].

3. CONVERGENCE WITH PROBABILITY ONE

As a first step towards ε-optimality, in this section we derive conditions on the algorithm such that the Markov process $\{p(k)\}$, $k \geqslant 0$, converges with probability one. To this end we begin by rewriting the consistency conditions (4) in a form more suitable for our analysis. Let

$$\begin{aligned} f_i^i[p] &= \alpha[i,p](1-p_i), \\ f_s^i[p] &= \beta[i,s,p]\,p_s, \\ \sum_{s \neq i} \beta[i,s,p]\,p_s &= \alpha[i,p](1-p_i) \end{aligned} \tag{C.1}$$

and

$$\begin{aligned} g_i^i[p] &= \gamma[i,p]\,p_i, \\ g_s^i[p] &= \delta[i,s,p](1-p_s), \\ \gamma[i,p]\,p_i &= \sum_{s \neq i} \delta[i,s,p](1-p_s), \end{aligned} \tag{C.2}$$

where $\alpha, \gamma : I \times S_M \to [0,1]$ and $\beta, \delta : I \times I \times S_M \to [0,1]$.

The basic rule that governs the choice of functions in the above form is that if a term is subtracted from p_j, then it is made proportional to p_j, and if a term is added to p_j, then it is made proportional to $1 - p_j$ irrespective of which coin is chosen for the toss and whether the toss results in success or failure.

REMARK 3. Our choice of the functions $g_s^i[\ ]$ is quite untraditional in the sense that in all most all the papers in mathematical psychology [18, 17, 13, 14] $g_s^i[p]$ is made proportional to p_s for all i and s, $s \neq i$. Also, in almost all the papers on learning automata [3–5, 7] $g_i^i[p]$ is made proportional to $1 - p_i$ and $g_s^i[p]$ is made proportional to p_s for all $s \neq i$. Because of this there is a disparity in the behavior of the algorithm (3) under success and failure. However, our present choice of functions $g_s^i[\ ]$ for all i and s given in (C.2) induce identical behavior of the algorithm (3) under success and failure.

The following theorem is immediate.

THEOREM 1. *Necessary and sufficient conditions for the learning algorithm* (3) *to have* V_M *as the only set of absorbing state are*:

(A.1) *For all* $P \in S_M - V_M$ *there exists* $1 \leqslant s \leqslant M$ *such that* $[\alpha[s,p]+\gamma[s,p]]p_s > 0$.

(A.2) *For all* $1 \leqslant s \leqslant M$, $\gamma[s,e_s] = 0$.

Proof. Refer to [6]. ■

A typical choice of functions in the algorithm (3) that satisfies the conditions (C.1)–(C.2) and (A.1)–(A.2) is given in the following example.

EXAMPLE 1. Let $0 < a_1, a_2 < 1$ and

$$\beta[i,s,p] = a_1(1-p_i)(1-p_s),$$

$$\alpha[i,p] = a_1 \sum_{s \neq i} p_s(1-p_s),$$

$$\delta[i,s,p] = a_2 p_i p_s^2 (1-p_i),$$

$$\gamma[i,p] = a_2(1-p_i) \sum_{s \neq i} p_s^2(1-p_s)$$

for all $i, s = 1,2,\ldots,M$.

REMARK 4. The demonstration of the symmetry in the properties of the algorithm (3) under success and failure alluded to in Remark 3 is evidenced by Theorem 1. Notice that V_M is the only set of absorbing states for (3) if $\alpha[i,p] \not\equiv 0$ and $\gamma[i,p] \equiv 0$, or $\alpha[i,p] \equiv 0$ and $\gamma[i,p] \not\equiv 0$, or both $\alpha[i,p]$ and $\gamma[i,p] \not\equiv 0$. This is in sharp contrast with the properties of the currently available absolutely expedient learning algorithm [3–5, 7], namely the reward-penalty and the reward-inaction algorithms have V_M as the only set of absorbing states, but the inaction-penalty algorithm does not. In fact, in all the inaction-penalty algorithms of the absolutely expedient type known so far [3], every state in δS_M is an absorbing state. Our modified definition of strong absolute expediency is in fact motivated by the existence of learning algorithms of the reward-penalty, reward-inaction, and inaction-penalty types, each with V_M as the only set of absorbing states.

REMARK 5. For some $1 \leqslant j \leqslant M$, if $p_j(k) = 0$, then it follows from (3) and (C.1)–(C.2) that $p_j(k^*) = 0$ for all $k^* \geqslant k$. In other words, during the learning process if $p(k)$ reaches the boundary $p_j = 0$ of the simplex S_M, then $p(k)$ will continue to remain in that boundary.

Henceforth in this paper we shall only be concerned with learning algorithms with V_M as the only set of absorbing states, that is, the algorithm (3) under conditions (A.1)–(A.2) of Theorem 1.

THEOREM 2. *Necessary and sufficient conditions for the learning algorithm* (3) *to be strongly absolutely expedient are*

$$\sum_{j \neq i} P_j \beta[i,j,p] = \sum_{j \neq i} P_j \beta[j,i,p] \tag{S.1}$$

and

$$\sum_{j \neq i} P_i(1-P_j)\delta[i,j,p] = \sum_{j \neq i} P_j(1-P_i)\delta[j,i,p] \tag{S.2}$$

for all $i = 1,2,\ldots,M$.

Sufficiency: Define $\delta x(k) = x(k+1) - x(k)$, and let

$$\Delta\eta(k) = E\big[\,\delta\eta(k)|p(k)\big] = \sum_{i=1}^{M} E\big[\,\delta P_i(k)|p(k)\big]\, d_i. \tag{11}$$

It can be seen by direct computation that

$$\begin{aligned} E\big[\,\delta p_i(k)|p(k)=p\big] = {} & P_i(1-p_i)d_i\alpha(i,p) - P_i^2 c_i \gamma[i,p] \\ & - \sum_{j \neq i} P_j P_i d_j \beta[j,i,p] \\ & + \sum_{j \neq i} P_j(1-p_i)c_j\delta[j,i,p]. \end{aligned} \tag{12}$$

Substituting (12) in (11), in view of (C.1) and (C.2), we obtain

$$\Delta\eta(k) = \Delta\eta_1(k) + \Delta\eta_2(k), \tag{13}$$

where

$$\Delta\eta_1(k) = \sum_{i=1}^{M} P_i d_i^2 \sum_{j \neq i} P_j \beta[i,j,p] - \sum_{i=1}^{M} P_i d_i \sum_{j \neq i} P_j d_j \beta[j,i,p] \tag{14}$$

and

$$\Delta\eta_2(k) = -\sum_{i=1}^{M} P_i d_i c_i \sum_{j \neq i} \delta[i,j,p](1-p_j)$$

$$+ \sum_{i=1}^{M} (1-P_i) d_i \sum_{j \neq i} P_j c_j \delta[j,i,p]. \tag{15}$$

Consider $\Delta\eta_1(k)$: It can be rewritten as

$$\Delta\eta_1(k) = \frac{b}{2},$$

where

$$b = \sum_{i=1}^{M} \sum_{j \neq i} \{ p_i p_j d_i^2 \beta[i,j,p] + p_i p_j d_i^2 \beta[i,j,p]$$

$$-2 p_i p_j d_i d_j \beta[j,i,p] \}.$$

Applying (S.1) to the second term within the curly braces, we obtain

$$b = \sum_{i=1}^{M} \sum_{j \neq i} \{ p_i p_j d_i^2 \beta[i,j,p] + p_i p_j d_i^2 \beta[j,i,p]$$

$$-2 p_i p_j d_i d_j \beta[j,i,p] \}$$

$$= P^T A P,$$

where $A = [A_{ij}]$, $A_{ii} = 0$, and

$$A_{ij} = d_i^2 \{ \beta[i,j,p] + \beta[j,i,p] \} - 2 d_i d_j \beta[j,i,p] \qquad \text{for} \quad i \neq j.$$

If we define a matrix $B = A + A^T = [B_{ij}]$, then

$$B_{ii} = 0$$

and

$$B_{ij} = \{ \beta[i,j,p] + \beta[j,i,p] \} (d_i - d_j)^2 \qquad \text{for} \quad i \neq j.$$

Since $b = \frac{1}{2}P^T BP$ and $B_{ij} \geqslant 0$, it readily follows that

$$\Delta\eta_1(k) = \frac{1}{4}\sum_{i=1}^{M}\sum_{j\neq i} p_i p_j (d_i - d_j)^2 \{\beta[i,j,p] + \beta[j,i,p]\}$$

$$\geqslant 0 \tag{16}$$

with equality holding only if $P \in V_M$.

Consider $\Delta\eta_2(k)$: It can be written as

$$\Delta\eta_2(k) = Z + g, \tag{17}$$

where

$$g = \sum_{i=1}^{M} p_i d_i^2 \sum_{j\neq i} (1-p_j)\delta[i,j,p]$$

$$- \sum_{i=1}^{M} (1-p_i) d_i \sum_{j\neq i} p_j d_j \delta[j,i,p] \tag{18}$$

and

$$Z = -\sum_{i=1}^{M} d_i \left\{ \sum_{j\neq i} p_i(1-p_j)\delta[i,j,p] - \sum_{j\neq i} p_j(1-p_i)\delta[j,i,p] \right\} \tag{19}$$

$$\equiv 0 \qquad \text{by the condition (S.2)} \tag{20}$$

If we define $y_i = 1 - p_i$ and $y = (y_1, y_2, \ldots, y_M)^T$, then g can be rewritten as

$$g = \frac{1}{2}\left\{ \sum_{i=1}^{M} p_i d_i^2 \sum_{j\neq i} y_i \delta[i,j,p] + \sum_{i=1}^{M} p_i d_i^2 \sum_{j\neq i} y_j \delta[i,j,p] \right.$$

$$\left. -2\sum_{i=1}^{M} y_i d_i \sum_{j\neq i} p_j d_j \delta[j,i,p] \right\}. \tag{21}$$

Applying (S.2) to the second term in the curly braces, we obtain

$$g = \tfrac{1}{2}\{ p^T E y + y^T F P \},$$

where

$$E = [E_{ij}], \qquad E_{ii} = 0, \qquad E_{ij} = d_i^2 \delta[i, j, p],$$

$$F = [F_{ij}], \qquad F_{ii} = 0, \qquad F_{ij} = (d_i^2 - 2d_i d_j)\delta[j, i, p]. \tag{22}$$

Since

$$g = \tfrac{1}{4}\{p^T(E + E^T)y + y^T(F + F^T)p\}$$

it follows after simplifications that

$$g = \frac{1}{4}\sum_{i=1}^{M}\sum_{j \neq i} p_i y_i (d_i - d_j)^2 \{\delta[i, j, p] + \delta[j, i, p]\}$$

$$\geqslant 0 \tag{23}$$

with equality holding only if $p \in V_M$. From (16) and (23) sufficiency follows.

Necessity: $\Delta\eta(k)$ can be represented as a quadratic and linear term in the vector D as follows:

$$\Delta\eta(k) = D^T A D + D^T B, \tag{24}$$

where $A = [A_{ij}]$ and $B = [B_1, B_2, \ldots B_M]^T$ with

$$\begin{aligned} A_{ii} &= P_i(1 - P_i)[i, p] + P_i^2 \gamma[i, p], \\ A_{ij} &= -\{P_i P_j \beta[j, i, p] + (1 - P_i) P_j \delta[j, i, p]\}, \end{aligned} \tag{25}$$

and

$$B_i = P_i^2 \delta[i, p] + (1 - P_i) \sum_{j \neq i} P_j \delta[j, i, p] \tag{26}$$

for all $i, j = 1, 2, \ldots, M$. From the definition of strong absolute expediency it follows that $\Delta\eta(k)$ attains its minimum value zero either when all d_i are equal or when $p \in V_M$. Since every member of V_M is absorbing on V_M, it is easily seen that $\Delta\eta(k)$ attains it minimum value for all admissible D-vectors. In the following we shall derive conditions for the minimum of $\Delta\eta(k)$ when $d_i = d$ for

all $i = 1,2,\ldots,M$, $0 < d < 1$. Necessary conditions for a minimum are obtained by setting the derivative of $\Delta\eta(k)$ (with respect to d_i) at the point $d_i = d$ for all $i = 1,2,\ldots,M$ equal to zero, that is,

$$\left.\frac{\partial \Delta\eta}{\partial d_i}\right|_{d_i = d} = 0 \qquad \text{for all} \quad i = 1,2,\ldots,M. \tag{27}$$

From (24), the equation (27) take the form

$$(A + A^T)d\mathbf{1} + B = \mathbf{0}, \tag{28}$$

where $\mathbf{1} = (1,1,\ldots,1)^T$ is an M-dimensional column vector of all ones. Rewriting (28), we get

$$dK_i + L_i = 0 \tag{29}$$

for all $i = 1,2,\ldots,M$ and all $0 < d < 1$, where

$$K_i = P_i(1 - P_i)\alpha[i,p] + P_i^2\gamma[i,p] - P_i \sum_{j \neq i} P_j\beta[j,i,p] - (1 - p_i) \sum_{j \neq i} P_i\delta[j,i,p]$$

and

$$L_i = P_i^2\gamma[i,p] + (1 - p_i) \sum_{j \neq i} P_j\delta[j,i,p].$$

(29) is true for all $0 < d < 1$ only if

$$L_i = 0 \text{ and } K_i = 0 \qquad \text{for all} \quad i = 1,2,\ldots,M.$$

And $L_i = 0$ leads to the condition

$$P_i^2\gamma[i,p] = (1 - P_i) \sum_{j \neq i} P_j\delta[j,i,p]. \tag{30}$$

Using (C.2), from (30) we obtain

$$P_i \sum_{j \neq i} (1 - P_j)\delta[i,j,p] = (1 - P_i) \sum_{j \neq i} P_j\delta[j,i,p], \tag{31}$$

which in fact is (S.2) Substituting (31) in $K_i = 0$, we obtain

$$P_i(1 - P_i)\alpha[i,p] = P_i \sum_{j \neq i} P_j\beta[j,i,p]. \tag{32}$$

Once again, using (C.1), we get

$$P_i \sum_{j \neq i} P_j \beta[i, j, p] = P_i \sum_{j \neq i} P_j \beta[j, i, p], \tag{33}$$

which is the same as (S.1). Hence the theorem.[1]

REMARK 6. It can be shown by routine computation that under the conditions (S.1)–(S.2) the following inequalities are true:

$$\Delta P_1(k) > 0 \quad \text{and} \quad \Delta P_M(k) < 0$$

for all $p(k) \in S_M - V_M$ and admissible D, where

$$\Delta P_j(k) = E\big[P_j(k+1) - P_j(k) | P(k) \big].$$

COROLLARY 1. *The Markov process* $\{p(k)\}$, $k \geq 0$, *as generated by the algorithm* (3) *under conditions* (C.1)–(C.2), (A.1)–(A.2), *and* (S.1)–(S.2) *converges to* V_M *with probability one.*

Proof. Since $\{p(k)\}$, $k \geq 0$, is a Markov process, from Theorem 2 we get

$$E\big[\delta\eta(k) | P(r) : 0 \leq r \leq k \big] = E\big[\delta\eta(k) | P(k) \big] \geq 0.$$

This in turn implies that $\{\eta(k)\}$, $k \geq 0$, is a submartingale [22]. By the martingale theorem, $\lim_{k \to \infty} \eta(k)$ exists and hence $\lim_{k \to \infty} p(k) = p^*$ exists with probability one. As $\delta\eta(k) = 0$ only on V_M, it follows that $P^* \in V_M$ with probability one.

4. ε-OPTIMALITY

In the previous section it was established that $p^* \in V_M$ with probability one. In this section we set out to quantify the distribution of P^*.

---

[1] The method of proof of the necessity part of Theorem 2 is the same as that used for proving absolute expediency of a more restricted class of algorithms in a forthcoming book [24].

To this end define

$$\Gamma_i(p) = \text{Prob}\left[P^* = e_i | P(0) = P\right] \tag{34}$$

for $i = 1,2,\ldots,M$. Notice $\sum_{i=1}^{M} \Gamma_i(p) = 1$ for all $p \in S_M$.

In view of Corollary 1 and (34), we obtain, for all $p(0) = p \in S_M^0$,

$$\lim_{k \to \infty} E[\eta(k)] = \sum_{i=1}^{M} \Gamma_i(p) d_i. \tag{35}$$

To compute $\Gamma_i(p)$ we need the following: Let $C[S_M]$ be the class of all continuous functions from S_M to the real line. If $f(\ ) \in C[S_M]$ define the operator U as follows

$$Uf(p) = E\left[f(p(k+1)) | p(k) = p\right]$$

Clearly the operator U is linear and positive [12].

DEFINITION 4. A function $f: S_M \to$ (real line) is called *superregular* (*regular*, *subregular*) if

$$f(p) \geqslant (=, \leqslant) Uf(p) \tag{36}$$

for all $p \in S_M$.

With these preliminaries, we now state two important propositions that lead an algorithm for quantifying $\Gamma_i(p)$.

PROPOSITION 1. *$\Gamma_i(p)$ is the only continuous solution of the functional equation*

$$U\Gamma_i(p) = \Gamma_i(p) \tag{37}$$

satisfying the boundary conditions

$$\Gamma_i(e_i) = 1 \quad \text{and} \quad \Gamma_i(e_j) = 0 \tag{38}$$

for all i, j, $i \neq j$.

The proof of this proposition is rather involved, and we refer the reader to Norman [12] for an elegant proof.

Notice that $\Gamma_i(p)$ satisfying (37) is a regular function by definition. This functional equation is extremely difficult to solve. Hence in the following we establish upper and lower bounds on $\Gamma_i(p)$.

PROPOSITION 2. *Let* $f_i(p) \in C[S_M]$ *be superregular (subregular) functions with* $f_i(e_i) = 1$ *and* $f_i(e_j) = 0$ *for all* $i, j, i \neq j$. *Then*

$$f_i(p) \geqslant (\leqslant) \Gamma_i(p). \tag{39}$$

Proof. Let $f_i(p)$ be a superregular function, that is,

$$f_i(p) \geqslant U f_i(p). \tag{40}$$

Since U is positive, we obtain

$$U f_i(p) \geqslant U^2 f_i(p) \geqslant \cdots \geqslant U^\infty f_i(p). \tag{41}$$

But since $\lim_{k \to \infty} p(k) = P^*$,

$$U^\infty f_i(p) = E\left[f_i(P^*) | P(0) = P \right] = \sum_{j=1}^{M} f_i(e_j) \Gamma_j(p) = \Gamma_i(p). \tag{42}$$

Combining (40), (41), and (42), we obtain

$$f_i(p) \geqslant \Gamma_i(p).$$

The result for subregular functions follow in a similar fashion. Hence the proposition.

Thus, if we can find two functions $h_i^1(p)$ and $h_i^2(p)$ which are super- and subregular functions respectively and satisfy the boundary conditions

$$\begin{aligned} &h_i^1(e_i) = h_i^2(e_i) = 1, \\ &h_i^1(e_j) = h_i^2(e_j) = 0 \qquad \text{for all} \quad j \neq i, \end{aligned} \tag{43}$$

then from Proposition 2 it will follow that

$$h_i^2(p) \leqslant \Gamma_i(p) \leqslant h_i^1(p). \tag{44}$$

Consider a function

$$\Psi_i[x_i, p] = e^{-x_i P_i/\theta}, \tag{45}$$

where $x_i > 0$ is a parameter. Recognizing the fact that

$$\phi_i[x_i, p] = \frac{1 - e^{-x_i P_i/\theta}}{1 - e^{-x_i/\theta}} \tag{46}$$

is subregular (superregular) whenever $\Psi_i[x_i, P]$ is subregular (superregular), it follows from Proposition 2 that

$$\phi_i[y_i P] \leqslant \Gamma_i(p) \leqslant \phi_i[x_i, p], \tag{47}$$

where y_i, x_i are two constants such that $\phi_i[y_i, p]$ and $\phi_i[x_i p]$ are sub- and superregular functions respectively. Note that function $\phi_i(x_i, p)$ depends on just one component of P and thus leads to conservative results. However, it gives rise to expressions which are easily manageable.

The problem of getting bounds on $\Gamma_i(p)$ now reduces to one of finding two positive constants y_i and z_i such that $\phi_i[y_i, p]$ is subregular and $\phi_i(x_i, P)$ is superregular. Further, from (35) and the inequality (1) it is clear that for ε-optimality we need to concentrate only (for the lower bound) on $\Gamma_1(p)$.

It can be seen that (dropping the subscript 1 from x, for convenience)

$$U\Psi_1[x, p] - \Psi_1[x, p] = -xF_1[x, p]\Psi_1[x, p], \tag{48}$$

where

$$\begin{aligned} F_1[x, p] = {} & p_1 d_1 \alpha[1, p](1 - p_1) V[-x\alpha[1, p](1 - p_1)] \\ & - p_1 c_1 \gamma[1, p] p_1 V[x\gamma[1, p] p_1] \\ & - \sum_{j \neq 1} p_j d_j \beta[j, 1, p] p_1 V[x\beta[j, 1, p] p_1] \\ & + \sum_{j \neq 1} p_j c_j \delta[j, 1, p](1 - p_1) V[-x\delta[j, 1, p](1 - p_1)] \end{aligned} \tag{49}$$

and

$$V[z] = \begin{cases} \dfrac{e^z - 1}{z} & \text{for } z \neq 0, \\ 1 & \text{for } z = 0. \end{cases} \tag{50}$$

Clearly, $F_1[x,p] \geqslant 0$ implies $\phi_1[x,p]$ subregular. Since $\alpha[i,p]$, $\gamma[i,p]$, $\delta[i,j,p]$ and $\beta[i,j,p]$ are all bounded above by unity and as the $V[\ ]$ is strictly monotonically increasing, it follows that

$$F_1[x,p] \geqslant 0$$

if

$$G[x,p] \triangleq \frac{V[-x(1-p_1)]}{V[xp_1]}$$

$$\geqslant \frac{P_1^2 c_1 \gamma[1,p] + \sum_{j \neq 1} p_1 p_j d_j \beta[j,1,p]}{\sum_{j \neq 1} p_j(1-p_1)\, c_j \delta[j,1,p] + p_1(1-p_1)\, d_1 \alpha[1,p]}. \tag{51}$$

It can be shown using the properties of the function $V[\ ]$ [4, 12] that

$$G[x,p] \geqslant \frac{1}{V[x]}. \tag{52}$$

Further, using (C.1)–(C.2) and (S.1)–(S.2), it can be seen that

$$e^* \triangleq \frac{c_1 A + d_2 B}{c_M A + d_1 B} \geqslant \frac{P_1^2 c_1 \gamma[1,p] + \sum_{j \neq 1} p_1 p_j d_j \beta[j,1,p]}{\sum_{j \neq 1} p_j(1-p_1)\, c_j \delta[j,1,p] + p_1(1-p_1)\, d_1 \alpha[1,p]}, \tag{53}$$

where

$$A = \sum_{j \neq 1} p_j(1-p_1)\ {}_j\delta[j,1,p],$$
$$B = \sum_{j \neq 1} p_1 p_j d_j \beta[j,1,p]. \tag{54}$$

In view of the inequality (1) it follows that $e^* < 1$. From (52) and (53)

$$F_1[x,p] \geqslant 0 \quad \text{if} \quad \frac{1}{V[x]} \geqslant e^*. \tag{55}$$

Since $e^* < 1$, $V[x] = 1/e^*$ has an unique solution $x = y > 0$ such that $\phi[y, p]$ is subregular.

REMARK 7. For the reward-inaction algorithm (see Remark 1) e^* reduces to d_2/d_1, and for the inaction-penalty algorithm e^* reduces to c_1/c_M. Thus in either of these special cases there exists an unique solution $x = y > 0$ for the equation $V[x] = 1/e^*$, and hence in both these special cases there exists a lower bound on $\Gamma_1(p)$. This is in sharp contrast with the existing results in the literature, wherein for the inaction-penalty (absolutely expedient) algorithms [4] no lower bound on $\Gamma_1(p)$ has been established. One of the reasons for this anomaly is that none of the currently available inaction-penalty (absolutely expedient) algorithms [4] have V_M as the only set of absorbing states.

Having established the lower bound on $\Gamma_1(p)$, we now state our main result.

THEOREM 3. *For every* $\varepsilon > 0$ *and* $p(0) = p \in S_M^0$ *there exists* $0 < \theta^* < 1$ *such that for all* $0 < \theta < \theta^*$, *the learning algorithm under the conditions* (C.1)–(C.2), (A.1)–(A.2), *and* (S.1)–(S.2) *is such that*

$$\lim_{k \to \infty} |E[\eta(k)] - d_1| < \varepsilon. \tag{56}$$

Proof. From Corollary 1 it follows that the limit on the left hand side of (56) exists. From (35)

$$\lim_{k \to \infty} E[\eta(k)] = \Gamma_1(p) d_1 + \sum_{j \neq 1} \Gamma_j(p) d_j. \tag{57}$$

That is,

$$\lim_{k \to \infty} |E[\eta(k) - d_1]| \leqslant (1 - \Gamma_1(p))|d_2 - d_1|. \tag{58}$$

Since $\phi_1[y, p] \leqslant \Gamma_1(p)$, combining this and (58), we have

$$\lim_{k \to \infty} |E[\eta(k)] - d_1| \leqslant (1 - \phi_1[y, p])|d_2 - d_1|. \tag{59}$$

From (46), for any $p \in S_M^0$, we know that

$$\lim_{\theta \to 0} \phi_1[x, p] = 1. \tag{60}$$

Combining (60) and (59), we see that for any $\delta > 0$ there exists $0 < \theta^* < 1$ such

that for all $0 < \theta < \theta^*$

$$\lim_{k \to \infty} |E[\eta(k)] - d_1| \leqslant |\delta(d_2 - d_1)|$$

The theorem follows by choosing

$$\delta = \frac{\varepsilon}{|d_2 - d_1|}.$$

5. SIMULATIONS

The algorithm (3) was simulated for the choice of the functions given in Example 1 above. The values of $\eta(k)$ averaged over 20 sample runs for various values of θ are given in Table 1.

6. CONCLUSIONS

A new class of strongly absolutely expedient learning algorithms, whose behavior under the action of success and failure is identical, has been introduced. Conditions for the ε-optimality of this class of learning algorithms have been derived.

The authors wish to thank an anonymous referee for suggesting various improvements.

TABLE 1

	$E[\eta(k)]$		
k	$\theta = 0.2$	$\theta = 0.1$	$\theta = 0.05$
0	0.5333	0.5333	0.5333
50	0.6379	0.5935	0.5668
100	0.6944	0.6510	0.6017
200	0.7505	0.7184	0.6598
500	0.7860	0.7693	0.7378
1000	0.7928	0.7865	0.7703
1500	0.7953	0.7913	0.7808
2000	0.7968	0.7938	0.7862
2500	0.7975	0.7952	0.7895
3000	0.7999	0.7996	0.7927

REFERENCES

1. K. S. Narendra and M. A. L. Thathachar, Learning automata—a survey, *IEEE Trans. Systems, Man, Cybernet.* 4:323–334 (1974).
2. K. S. Narendra and S. Lakshmivarahan, Learning automata—a critique, *J. Cybernet. Inform. Sci.* (special issue on learning automata), 1:53–66 (1978).
3. S. Lakshmivarahan and M. A. L. Thathachar, Absolutely expedient learning algorithms for stochastic automata, *IEEE Trans. Systems, Man, Cybernet.* 3:281–286 (1973).
4. S. Lakshmivarahan and M. A. L. Thathachar, Bounds on the probability of convergence of learning automata, *IEEE Trans. Systems, Man, Cybernet.*, 6:756–763 (1976).
5. H. Aso and M. Kimura, Absolute expediency of learning automata, *Inform. Sci.*, 17:91–112 (1979).
6. U. Herkenrath, D. Kalin, and S. Lakshmivarahan, On a general class of absorbing barrier learning algorithms, School of EECS Technical Report 8001, Univ. of Oklahoma, Mar. 1980; *Inform. Sci.* 23: (1981).
7. Y. Sawaragi and N. Baba, Two ε-optimal non-linear reinforcement schemes for stochastic automata, *IEEE Trans. Systems, Man, Cybernet.* 4:126–131 (1974).
8. K. S. Fu, Stochastic automata models for learning systems, in *Computers and Information Sciences II* (J. T. Tou, Ed.), Academic, 1967.
9. G. J. McMurtry and K. S. Fu, A variable structure automaton used in multimodal search technique, *IEEE Trans. Automat. Control* 11:379–387 (1966).
10. V. I. Varshavskii and I. P. Vorontsova, On the behaviour of stochastic automata with variable structure, *Automat. Remote Control* 24:327–333 (1963).
11. M. L. Tsetlin, *Automaton Theory and Modelling of Biological Systems*, Academic, 1973.
12. M. F. Norman, On a linear model with two absorbing barriers, *J. Math. Psych.* 5:225–241 (1968).
13. M. F. Norman, *Markov Processes and Learning Models*, Academic, New York, 1972.
14. M. Iosifescu and R. Theodorescu, *Random Processes and Learning*, Springer, 1969.
15. S. Lakshmivarahan, ε-optimal learning algorithms—Nonabsorbing barrier type, Technical Report EECS 7901, School of EECS, Univ. of Oklahoma, Feb. 1979.
16. I. H. Witten, The apparent conflict between estimation and control—a survey of two armed-bandit problem, *J. Franklin Inst.* 301:161–189 (1976).
17. B. Chandrasekaran and D. W. C. Shen, On expediency and convergence in variable structure automata, *IEEE Trans. Systems Sci. Cybernet.* 4:52–60 (1968).
18. R. R. Bush and F. Mosteller, *Stochastic Models for Learning*, Wiley, New York, 1958.
19. R. Viswanathan and K. S. Narendra. Expedient and optimal variable structure stochastic automata, Dunharm Lab., TR-37, Yale Univ., 1970.
20. I. J. Shapiro and K. S. Narendra, "Use of stochastic automata for parameter self-optimisation with multimodal performance criteria, *IEEE Trans. Systems, Man, Cybernet.* 5:352–360 (1969).
21. S. Lakshmivarahan, *Learning Algorithms: Theory and Applications*, Springer, New York, 1981.
22. J. L. Doob, *Stochastic Processes*, Wiley, 1955.
23. M. R. Meybodi and S. Lakshmivarahan, On a class of learning algorithms with symmetric behaviour under success and failure, in *Proceedings of the Conference on Mathematical Learning Models—Theory and Applications*, 3–7 May, Springer Verlag Lecture Notes in Statistics, 1982.
24. K. S. Narendra and M. A. L. Thathachar, *Learning Automata*, to appear.

Received October 1981; revised August 1982