NOT FOR QUOTATION
WITHOUT PERMISSION
OF THE AUTHOR

LEARNING BEHAVIORS OF STOCHASTIC AUTOMATA AND SOME APPLICATIONS

Norio Baba

November 1983
WP-83-119

Working Papers are interim reports on work of the International Institute for Applied Systems Analysis and have received only limited review. Views or opinions expressed herein do not necessarily represent those of the Institute or of its National Member Organizations.

INTERNATIONAL INSTITUTE FOR APPLIED SYSTEMS ANALYSIS
2361 Laxenburg, Austria

PREFACE

It is known that stochastic automata can be applied to describe the behavior of a decision maker or manager in the condition of uncertainty. This paper discusses learning behaviors of stochastic automata under unknown nonstationary multi-teacher environment. The consistency of sequential decision making procedures is proved under some mild conditions.

V. Fedorov

ACKNOWLEDGMENTS

The author would like to thank Prof. Fedorov, Prof. Holling, Prof. Kindler, Prof. Sawaragi, Prof. Shoman, Prof. Soeda, Dr. Staley, Prof. Walters, Prof. Wets, Prof. Wierzbicki, and Prof. Yashin for their kind advice and encouragement.

CONTENTS

LEARNING BEHAVIORS OF STOCHASTIC AUTOMATA AND SOME APPLICATIONS

Norio Baba

INTRODUCTION

The concept of learning automata operating in an unknown random environment was first introduced by Tsetlin (1961). He studied the learning behaviors of deterministic automata and showed that they are asymptotically optimal under some conditions. Later, Varshavskii and Vorontsova (1963) found that stochastic automata also have learning properties. Since then, the learning behaviors of stochastic automata have been studied extensively by many researchers. Chandrasekaran and Shen (1968), Norman (1968; 1972), Lakshmivarahan and Thathachar (1973), Narendra and Thathachar (1974), and others, have contributed fruitful results to the literature of learning automata.

Despite active research in this field, almost all research so far has dealt with learning behaviors of a single automaton operating in a stationary single-teacher environment, although Koditschek and Narendra (1977) considered the learning behavior of fixed-structure automata operating in a stationary multi-teacher environment. Thathachar and Bhakthavathsalam (1978) then studied variable-structure stochastic automata operating in two distinct teacher environments. Recently, Baba (1983) studied the learning behaviors of variable-structure stochastic automata under the general n-teacher environment. He proposed the GAE reinforcement scheme as a learning algorithm and proved that this reinforcement scheme has good learning properties such as ε-optimality and absolute expediency in the general n-teacher environment.

In this paper, we consider learning behaviors of variable-structure stochastic automata operating in a nonstationary multi-teacher environment from which stochastic automata receive responses having an arbitrary number between 0 and 1. As a generalized form of the GAE reinforcement scheme, we propose the MGAE scheme and show that this scheme ensures ε-optimality in the nonstationary multi-teacher environment of an S-model. We also consider the parameter self-optimization problem with noise-corrupted, multi-objective functions by stochastic automata.

Since the theory of the learning behavior of stochastic automata operating in the NMT environment has been developed only recently, its application to real problems has not been discussed in the literature. However, the author believes that it could be applied to the problems where one input elicits multi-responses from multi-criteria environments. In the following, we shall suggest two applications:

Commercial Game

Suppose that n players $(A_1,\ldots,A_n)$ are taking part in a game in which they wish to open a store somewhere in r regions $(B_1,\ldots,B_r)$. The mth player (A_m) will choose the region B_k with a probability p_{mk} $(m=1,\ldots,n;\ k=1,\ldots,r)$. It is assumed that we can not obtain any information about these probabilities. However, if a player is to succeed, he must avoid regions containing a lot of other players. The MGAE reinforcement scheme, which will be proposed in this paper, can be used to find an appropriate region where there is a minimum of overlapping. The learning behavior of automata using the MGAE scheme in various commercial games has been simulated by computer and results indicating the effectiveness of the scheme have been obtained.

Fishing

Suppose that there are r sea-areas in which a group of ships $(S_1,...,S_n)$ must catch fish. The learning behaviors of stochastic automata under multi-teacher environments can also be applied to find an appropriate sea-area. In this case, n ships and r sea-areas become n teachers and the r states of the stochastic automaton, respectively. If the numbers (or volume) of the catches of the ith ship S_1 are low, S_i emits a penalty response. On the contrary, if great numbers of catches have been obtained, then S_i emits a reward response. Depending upon the n responses from n teachers, the stochastic automaton changes its state probability vector.

BASIC MODEL OF A LEARNING AUTOMATON OPERATING IN AN UNKNOWN ENVIRONMENT

The learning behaviors of a variable-structure stochastic automaton operating in an unknown random environment have been discussed extensively under the model shown in Figure 1. First, let us briefly explain the learning mechanism of the stochastic automaton A under the unknown random environment (teacher environment) $R(C_1,...,C_r)$. Then, we will explain the basic norms of the learning behaviors of the stochastic automaton A.

The stochastic automaton A is defined by the sextuple $\{S,W,Y,g,P(t),T\}$. S denotes the set of two inputs (0.1), where 0 indicates the reward response from $R(C_1,...,C_r)$ and 1 indicates the penalty response. (If the set S consists of only two elements 0 and 1, the environment is said to be a P-model. When the input into A assumes a finite number of values in the closed interval $[0,1]$, it is said to be a Q-model. An S-model is one in which the input into A takes an arbitrary number in the closed line segment $[0,1]$. In the next section, we will deal with the S-model envirnment.) W denotes the set of r internal states $(w_1,...,w_r)$. Y denotes the set of r outputs $(y_1,...,y_r)$. g denotes the output function $y(t)=g[w(t)]$, that is, one to one deterministic mapping. $P(t)$ denotes the probability vector $[p_1(t),...,p_r(t)]'$ at time t, and its ith component $p_i(t)$ indicates the probability with which the ith state w_i is chosen at time $t\,(i=1,...,r)$:

$$p_1(0) = \cdots = p_r(0) = \frac{1}{r}, \qquad \sum_{i=1}^{r} p_i(t) = 1$$

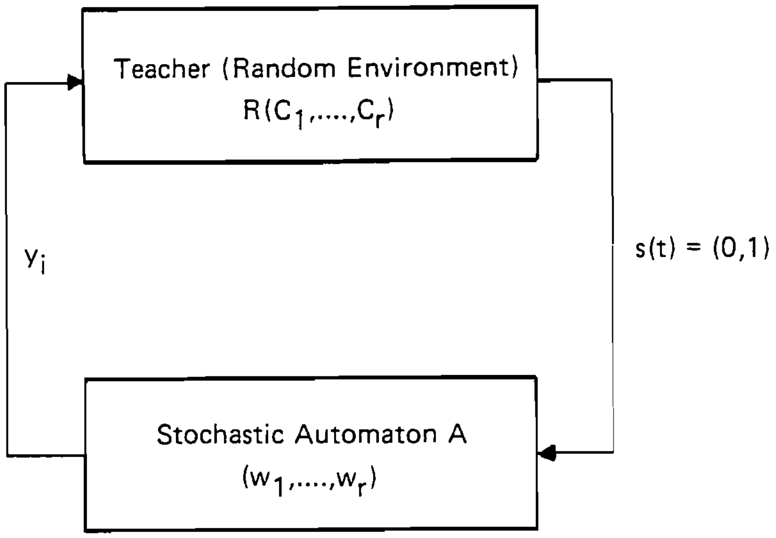

Figure 1. Basic model of a learning automaton operating in an unknown random environment.

T denotes the reinforcement scheme which generates $P(t+1)$ from $P(t)$.

Suppose that the state w_i is chosen at time t. Then, the stochastic automaton A performs action y_i on the random environment $R(C_1,\ldots,C_r)$. In response to the action y_i, the environment emits output $s(t)=1$ (penalty) with probability C_i and output $s(t)=0$ (reward) with probability $1-C_i (i=1,\ldots,r)$. If all of the $C_i (i=1,\ldots,r)$ are constant, the random environment $R(C_1,\ldots,C_r)$ is said to be a stationary random environment. (The term "single teacher environment" is also used instead of the term "random environment.") On the other hand, if $C_i (i=1,\ldots,r)$ are not constant, it is said to be a nonstationary random environment. Depending upon the action of the stochastic automaton A and the environmental response to it, the reinforcement scheme T changes the probability vector $P(t)$ to $P(t+1)$.

The values of $C_i (i=1,\ldots,r)$ are not known *a priori*. Therefore, it is necessary to reduce the average penalty,

$$M(t) = \sum_{i=1}^{r} p_i(t) C_i \tag{1}$$

by selecting an appropriate reinforcement scheme. To judge the effectiveness of a learning automaton operating in a stationary random environment $R(C_1,\ldots,C_r)$, various performance measures have been set up. (See Chandrasekaran and Shen 1968; Lakshmivarahan and Thathachar 1973; Narendra and Thathachar 1974.)

DEFINITION 1. *A reinforcement scheme is said to be expedient if*

$$\lim_{t\to\infty} E\{M(t)\} < \{\frac{1}{r}\sum_{i=1}^{r} C_i\} \tag{2}$$

($E\{\cdot\}$ is the mathematical expectation.)

DEFINITION 2. *A reinforcement scheme is said to be optimal if*

$$\lim_{t\to\infty} E\{p_\alpha(t)\} = 1 \tag{3}$$

where $C_\alpha = \min_i \{C_i\}$

DEFINITION 3. *A reinforcement scheme is said to be ε-optimal if*

$$\lim_{\vartheta \to 0} \lim_{t \to \infty} E\{p_\alpha(t)\} = 1$$

where ϑ is a parameter included in the reinforcement scheme.

DEFINITION 4. *A reinforcement scheme is said to be absolutely expedient if*

$$E\{M(t+1)/P(t)\} < M(t) \tag{4}$$

for all t, all $p_i(t) \in (0,1)$ $(i=1,\ldots,r)$, and all possible values of C_i $(i=1,\ldots,r)$. ($E\{M(t+1)/P(t)\}$ is the conditional expectation.)

Remarks.

(a) The definition of ε-optimality can also be described by using $M(t)$.

(b) In Definition 4, the trivial case in which all the values of $C_i(i=1,\ldots,r)$ are equal is precluded.

The learning behaviors of a variable-structure stochastic automaton operating in the stationary random environment $R(C_1,\ldots,C_r)$ have been extensively studied by many researchers. Norman (1968) proved that the L_{R-I} scheme ensures ε-optimality in the two state case. Sawaragi and Baba (1973) showed that this property also holds in the general r-state case. Lakshmivarahan and Thathachar (1973) introduced the concept of absolutely expedient learning algorithms.

Remark.

(c) L_{R-I} scheme (Reward-Inaction scheme)

Assume $y(t)=y_i$.

If $s(t)=0$, then

$$p_i(t+1) = (1-\vartheta)p_i(t)+\vartheta \qquad 0<\vartheta<1$$

$$p_j(t+1) = (1-\vartheta)p_j(t) \qquad (j \neq i)$$

If $S(t)=1$, then

$$p_m(t+1) = p_m(t) \qquad (m=1,\dots,r)$$

Compared with the great number of studies related to the behavior of learning automata in a stationary environment, only a few and specialized results have been obtained concerning those in a nonstationary environment. Babà and Sawaragi (1975) considered the nonstationary random environment which has the property that

$$C_\alpha(t,\omega)+\delta_1 < C_j(t,\omega)$$

(holds for some α, some $\delta_1>0$, all $j(\neq\alpha)$, all t, and all ω; ω is a point of a basic ω-space Ω.)

They showed that the L_{R-I} scheme ensures ε-optimality under the above environment. Recently, Srikantakumar and Narendra (1982) studied the learning behaviors of stochastic automata under the following nonstationary random environment:

(i) $C_i[P(n)]$ $(i=1,\dots,r;\ n=0,\dots)$ are continuous functions of $p_i(i=1,\dots,r)$

(ii) $\dfrac{\partial C_i}{\partial p_i} > 0$ for $\forall i$

(iii) $\dfrac{\partial C_i}{\partial p_i} \gg \dfrac{\partial C_i}{\partial p_j}$ for $\forall i,j$ $(i\neq j)$

This work has a very interesting application in the area of telephone network routing.

LEARNING AUTOMATON MODEL UNDER THE NONSTATIONARY MULTI-TEACHER ENVIRONMENT

In this section, we generalize the model given in Figure 1 and discuss the learning behaviors of the variable-structure stochastic automaton B in the nonstationary multi-teacher environment (NMT) as illustrated in Figure 2.

The stochastic automaton B is defined by the set $\{S, W, Y, g, P(t), T\}$. S is the set of inputs $(S_1^i, \ldots, S_n^i)$ where $S_j^i (j=1,\ldots,n)$ is the response from the jth teacher $R_j (j=1,\ldots,n)$ and the value of S_j^i can be an arbitrary number in the closed line segment $[0,1]$. (We are dealing with an S-model multi-teacher environment.) In this model, the definitions of $W, Y, g, P(t)$, and T are the same as in the last section.

Assume now that the state w_i is chosen at time t. Then, the stochastic automaton B performs action y_i on the nonstationary multi-teacher environment (NMT). In response to y_i, the jth teacher R_j emits output S_j^i. In this section, we shall deal with the case in which S_j^i is a function of t and ω. ($\omega \varepsilon \Omega$; Ω is the basic ω-space of the probability measure space $(\Omega, \tilde{B}, \mu)$, and $\tilde{B}$ is the smallest Borel field including $\bigcup_{t=0}^{\infty} F_t$, where $F_t = \sigma[(P(0),\ldots,P(t), C(0),\ldots,C(t)].$) Consequently, from now on we shall use the notation $S_j^i(t,\omega)$ to represent the input into the stochastic automaton B.

Depending upon the action y_i and the n responses $S_1^i(t,\omega),\ldots,S_n^i(t,\omega)$ from n teachers $R_1,\ldots,R_n$, the stochastic automaton B changes the probability vector $P(t)$ by the reinforcement scheme T.

The nonstationary multi-teacher environment (NMT) considered in this paper has the property that the relation

$$\int_0^1 sdF_{\alpha,t}(s) + \frac{\delta}{n} < \int_0^1 sdF_{j,t}\,(s) \tag{5}$$

where $F_{i,t}(s)$ $(i=1,\ldots,r)$ is the distribution function of $\dfrac{s_1^i(t,\omega) + \cdots + s_n^i(t,\omega)}{n}$, holds for some state w_α, some $\delta>0$, all time t, all $j\,(\ddagger\alpha)$, and all $\omega(\in\Omega)$.

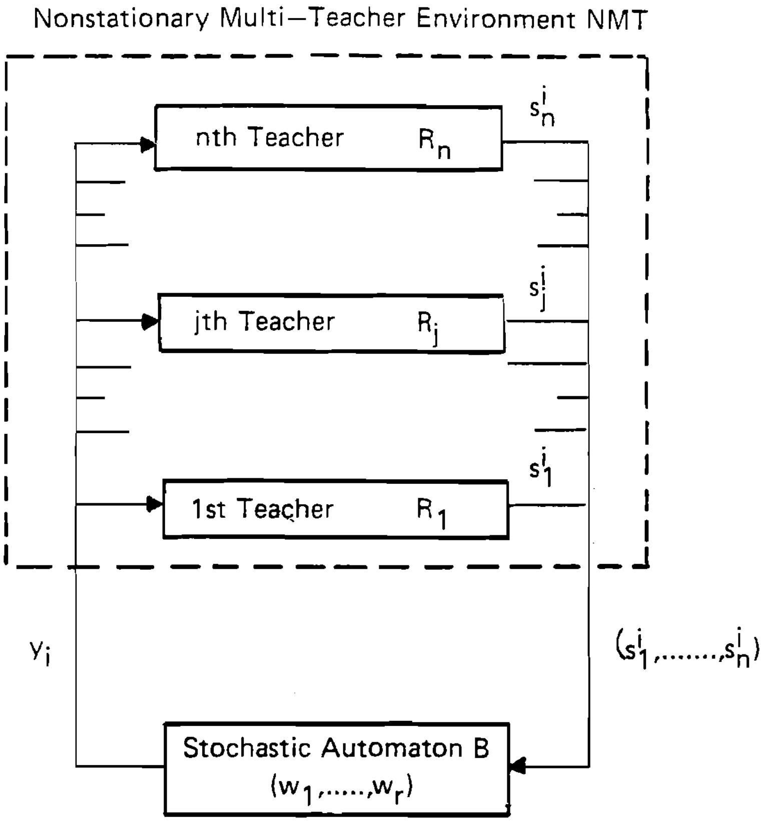

Figure 2. Stochastic automaton *B* operating in the nonstationary multi-teacher environment (NMT).

The objective of the stochastic automaton B is to reduce $E\{\sum_{j=1}^{n} S_j^i(t,\omega)\}$, the expectation of the sum of the penalty strengths. Therefore, condition (5) means that the αth action y_α is the best among r actions $y_1,\dots,y_r$ since y_α receives the least sum of the penalty strengths in the sense of mathematical expectation.

Before we proceed to introduce the norms of learning behaviors of stochastic automata under an NMT environment, we will explain several basic norms of the learning behaviors of stochastic automata under a stationary multi-teacher environment of a P-model. Baba (1983) discussed the learning behaviors of stochastic automata operating in the general stationary n-teacher environment in which there exists a βth state w_β such that

$$C_1^\beta + \dots + C_n^\beta < C_1^i + \dots + C_n^i \tag{6}$$

for all $1 \leq i \leq r \quad (i \ddagger \beta)$

He gave the following definitions:

DEFINITION 1. *The average weighted reward in the n-teacher environment $\overline{W}(t)$ is defined as follows:*

$$\overline{W}(t) = \sum_{i=1}^{r} [p_i(t)\{\sum_{j=1}^{n} jD_{n,j}^i\}] \tag{7}$$

where $D_{n,j}^i$ is the probability that j teachers approve of the ith action y_i of the stochastic automaton B. $(j=1,\dots,n)$

DEFINITION 2. *The stochastic automaton B is said to be "absolutely expedient in the general n-teacher environment" if*

$$E\{\overline{W}(t+1)/P(t)\} > \overline{W}(t) \tag{8}$$

for all t, all $p_i(t) \in (0,1)$, $i=1,\dots,r$, and all $C_k^i \in (0,1)$, $i=1,\dots,r$; $k=1,\dots,n$.

DEFINITION 3. *The stochastic automaton B is said to be "expedient in the general n-teacher environment" if*

$$\lim_{t\to\infty} E\{\overline{W}(t)\} > \overline{W}_0 \tag{9}$$

where $\overline{W}_0 = \sum_{i=1}^{r} \frac{1}{r}\{\sum_{j=1}^{n} jD_{n,j}^{i}\}$.

DEFINITION 4. *The stochastic automaton B is said to be "optimal in the general n-teacher environment" if*

$$\lim_{t\to\infty} p_\beta(t) = 1 \tag{10}$$

with probability 1.

DEFINITION 5. *The stochastic automaton B is said to be ε-optimal in the general n-teacher environment" if one can choose parameter ϑ included in the reinforcement scheme of stochastic automaton B such that*

$$\lim_{\vartheta\to 0}\lim_{t\to\infty} E\{p_\beta(t)\} = 1 \tag{11}$$

Baba proposed the GAE reinforcement scheme and proved that it ensures ε-optimality and absolute expediency in the general n-teacher environment.

By analogy from Definitions 4 and 5 given above, we can give the following definitions concerning learning norms of stochastic automata under nonstationary multi-teacher environment NMT satisfying the condition (5):

DEFINITION 6. *The stochastic automaton B is said to be optimal in NMT if*

$$\lim_{t\to\infty} p_\alpha(t) = 1 \tag{12}$$

with probability 1.

DEFINITION 7. *The stochastic automaton B is said to be ε-optimal in NMT if one can choose a parameter ϑ included in the reinforcement scheme T of the stochastic automaton B such that the following equality holds:*

$$\lim_{\vartheta \to 0} \lim_{t \to \infty} E\{p_\alpha(t)\} = 1 \tag{13}$$

On the other hand, the extensions of Definitions 2 and 3 can not be easily given. Presumably, we need a different interpretation.

ε-OPTIMAL REINFORCEMENT SCHEME UNDER THE NONSTATIONARY MULTI-TEACHER ENVIRONMENT

The GAE reinforcement scheme (Baba 1983) has been introduced as a class of learning algorithms of stochastic automata operating in a multi-teacher environment which emits 0 (reward) or 1 (penalty) responses. This scheme can not be applied to the S-model environment in which teachers emit arbitrary responses between 0 and 1.

In the following, let us propose the MGAE scheme which can be used for the S-model environment.

MGAE SCHEME:

Suppose that $y(t)=y_i$ and the responses from NMT are $(s_1^i, s_2^i, \dots, s_n^i)$. $(s_j^i (j=1,\dots,n)$ means the response from the jth teacher.) Then,

$$p_i(t+1) = p_i(t) + \left[\frac{s_1^i + \dots + s_n^i}{n}\right]\{\sum_{j \neq i}^{r} \varphi_j[P(t)]\} - \left[1 - \frac{s_1^i + \dots + s_n^i}{n}\right]\{\sum_{j \neq i}^{r} \psi_j[P(t)]\} \tag{14}$$

$$p_j(t+1) = p_j(t) - \left[\frac{s_1^i + \dots + s_n^i}{n}\right]\{\varphi_j[P(t)]\} + \left[1 - \frac{s_1^i + \dots + s_n^i}{n}\right]\{\psi_j[P(t)]\} \qquad (j \neq i) \tag{15}$$

where $\varphi_i, \psi_i (i=1,\dots,r)$ satisfy the following relations.

$$\frac{\varphi_1[P(t)]}{p_1(t)} = \frac{\varphi_2[P(t)]}{p_2(t)} = \cdots = \frac{\varphi_r[P(t)]}{p_r(t)} = \lambda[P(t)] \tag{16}$$

$$\frac{\psi_1[P(t)]}{p_1(t)} = \frac{\psi_2[P(t)]}{p_2(t)} = \cdots = \frac{\psi_r[P(t)]}{p_r(t)} = \mu[P(t)] \tag{17}$$

$$p_j(t) + \psi_j[P(t)] > 0 \tag{18}$$

$$p_i(t) + \sum_{j \neq i}^{r} \varphi_j[P(t)] > 0$$

$$p_j(t) - \varphi_j[P(t)] < 1 \qquad (j=1,\dots,r \quad i=1\dots,r)$$

As to the learning performance of the MGAE reinforcement scheme, the following theorem can be obtained.

THEOREM 1. *Suppose that* $\lambda[P(t)] = \vartheta\{\bar{\lambda}[P(t)]\}$ $(\vartheta>0)$ *(19) and* $\mu[P(t)] = \vartheta\{\bar{\mu}[P(t)]$ *(20), where* $\bar{\lambda}[P(t)]$ *and* $\bar{\mu}[P(t)]$ *are bounded functions which satisfy the following conditions:* $\bar{\lambda}[P(t)] \leq 0$ *(21),* $\bar{\mu}[P(t)] \leq 0$ *(22), and* $\bar{\lambda}[P(t)] + \bar{\mu}[P(t)] < 0$ *(23), for all* $P(t)$ *and* t.

Then, the stochastic automaton B with the MGAE reinforcement scheme is ε-optimal under the nonstationary multi-teacher environment NMT satisfying condition (5).

Since the proof of the above theorem is rather lengthy, we will begin by deriving several important lemmas.

LEMMA 1. *Suppose that all of the assumptions of the above theorem hold. Then, the MGAE reinforcement scheme has the following learning performance under the NMT environment satisfying condition (5):*

$$E\{p_\alpha(t+1)/P(t)\} \geq p_\alpha(t)$$

Proof. For notational convenience, let us abbreviate time t and probability vector $P(t)$ as follows:

$$p_i = p_i(t), \qquad \varphi_i = \varphi_i[P(t)], \qquad \psi_i = \psi_i[P(t)],$$

$$\lambda = \lambda[P(t)], \qquad u = u[P(t)]. \qquad (i=1,\ldots,r)$$

Let $F_{i,t}(\xi)$ be the distribution function of

$$\frac{s_1^i(t,\omega)+\ldots+s_n^i(t,\omega)}{n} \qquad (i=1,\ldots,r) \tag{24}$$

Then, the conditional expectation $E\{p_\alpha(t+1)/P(t)\}$ can be calculated as follows:

$$E[p_\alpha(t+1)/P(t)] = p_\alpha \int_0^1 [p_\alpha + \xi(\sum_{j\neq\alpha}^r \varphi_j) - (1-\xi)(\sum_{j\neq\alpha}^r \psi_j)]\, dF_{\alpha,t}(\xi)$$

$$+ \sum_{j\neq\alpha}^r p_j \int_0^1 [p_\alpha - \xi(\varphi_\alpha) + (1-\xi)\psi_\alpha]\, dF_{j,t}(\xi)$$

$$= p_\alpha - p_\alpha(\sum_{j\neq\alpha}^r \psi_j) + p_\alpha[\sum_{j\neq\alpha}^r (\varphi_j+\psi_j)][\int_0^1 \xi\, dF_{\alpha,t}(\xi)]$$

$$+ (1-p_\alpha)\psi_\alpha - (\varphi_\alpha+\psi_\alpha)[\sum_{j\neq\alpha}^{r} p_j \int_0^1 \xi dF_{j,t}(\xi)] \qquad (25)$$

Let

$$C_k(t) = [\int_0^1 \xi dF_{k,t}(\xi)] \qquad (k=1,\dots,r) \qquad (26)$$

Then, using the relations (16) and (17), the above equality can be represented as:

$$E[p_\alpha(t+1)/ P(t)] = p_\alpha + p_\alpha(1-p_\alpha)[\lambda+\mu]C_\alpha(t) \qquad (27)$$

$$- [\lambda+\mu][\sum_{j\neq\alpha}^{r} p_j C_j(t)]p_\alpha$$

From the definition of the distribution function $F_{k,t}(\xi)$, and from condition (5),

$$C_\alpha(t) + \frac{\delta}{n} < C_j(t) \qquad \text{for } \mathit{all}\ j(j\neq\alpha,\ 1\leq j\leq r) \text{ and } \omega \qquad (28)$$

Let

$$C_\beta(t) = min[C_{k_1}(t),\dots,C_{k_{r-1}}(t)] \qquad (k_1,\dots,k_{r-1}\neq\alpha) \qquad (29)$$

Then, from the relations (19) ~ (29), we can get

$$E[p_\alpha(t+1)]/ P(t)] \geq p_\alpha(t) + [\lambda+\mu](1-p_\alpha)p_\alpha[C_\alpha(t) - C_\beta(t)] \geq p_\alpha(t)$$

$$[C_\alpha(t) - C_\beta(t) < 0 \quad \text{and} \quad \lambda+\mu < 0] \qquad (30)$$

Remark. (30) is the Semi-Martingale Inequality (Doob 1953). From this inequality, we can get $E[p_\alpha(t+1)] \geq E[p_\alpha(t)]$ for all t. This means that the mathematical expectation $p_\alpha(t)$ increases monotonously with time t.

LEMMA 2. *Suppose that all of the assumptions of the theorem hold. Let*

$$h_{x,\vartheta}(p) = \frac{\exp(xp/\vartheta)-1}{\exp(x/\vartheta)-1} \qquad (x>0) \tag{31}$$

$$p_\alpha'(t) = 1-p_\alpha(t) \tag{32}$$

Then, there exists some positive constant z which satisfies the inequality

$$E\{h_{z,\vartheta}[p_\alpha'(t+1)]/P(t)\} \le h_{z,\vartheta}[p_\alpha'(t)] \qquad \text{for } \textit{all } t \text{ and } P(t)$$

Proof. The conditional expectation $E\{h_{x,\vartheta}[p_\alpha'(t+1)]/P(t)\}$ can be calculated as follows:

$$\begin{aligned} E\{h_{x,\vartheta}[p_\alpha'(t+1)]/P(t)\} = J[-1 & \\ &+ \exp\frac{xp_\alpha'}{\vartheta}\{p_\alpha\int_0^1 \exp\left[-xp_\alpha'(\xi\bar{\lambda}-(1-\xi)\bar{\mu})\right] dF_{\alpha,t}(\xi) \\ &+ \sum_{j\neq\alpha}^{r} p_j \int_0^1 \exp[xp_\alpha(\xi\bar{\lambda}-(1-\xi)\bar{\mu})dF_{j,t}(\xi)\}] \end{aligned} \tag{33}$$

where

$$J = \frac{1}{\exp\frac{x}{\vartheta}-1}, \qquad p_\alpha'=1-p_\alpha(t), \qquad \text{and} \quad p_\alpha=p_\alpha(t)$$

Assume that

$$|\bar{\lambda}+\bar{\mu}| < 0_1 \qquad (0_1: \textit{positive constant}) \tag{34}$$

Then, by using Taylor's expansion theorem, the following two inequalities can be obtained:

$$\begin{aligned} exp[-xp_\alpha'(\xi\bar{\lambda}-(1-\xi)\bar{\mu}) \le 1 & -xp_\alpha'(\xi\bar{\lambda}-(1-\xi)\bar{\mu}) \\ &+ 20_1|\bar{\lambda}+\bar{\mu}|x^2p_\alpha'[exp(20_1x)] \end{aligned} \tag{35}$$

$$exp\,[xp_{\alpha}(\xi\bar{\lambda}-(1-\xi)\bar{\mu})] \le 1 \ +xp_{\alpha}(\xi\bar{\lambda}-(1-\xi)\bar{\mu}) \\ + 2O_1|\bar{\lambda}+\bar{\mu}|x^2p_{\alpha}[exp\,(2O_1x)] \tag{36}$$

From (33), (35), and (36), we can get

$$E\{h_{x,\vartheta}[p_{\alpha}'(t+1)]/P(t)\} \le -J \ + J[\exp\frac{xp_{\alpha}'}{\vartheta}]\{1 \\ -p_{\alpha}|\bar{\lambda}+\bar{\mu}|x[(p_{\alpha}'C_{\alpha}(t) \\ -\sum_{j\neq\alpha}^{r} p_j C_j(t)) + f_1(x,P)]\} \tag{37}$$

where

$$f_1(x,P) = -4x\,O_1p_{\alpha}'[exp\,(2O_1x)] \tag{38}$$

From (28),

$$E\{h_{x,\vartheta}[p_{\alpha}'(t+1)]/P(t)\} \le -J \ + J(exp\frac{xp_{\alpha}'}{\vartheta})[1 \\ -p_{\alpha}p_{\alpha}'|\bar{\lambda}+\bar{\mu}|x(\frac{\delta}{n} \\ -4x\,O_1\exp(2O_1x))] \tag{39}$$

In the above equality (39), $\lim_{x\to 0} 4x\,O_1\exp(2O_1x) = 0$, $p_{\alpha}p_{\alpha}'|\bar{\lambda}+\bar{\mu}| \ge 0$, and $\frac{\delta}{n}$ is a positive constant. Hence, there exists some positive constant z which satisfies the inequality

$$E\{h_{z,\vartheta}[p_{\alpha}'(t+1)]/P(t)\} \le h_{z,\vartheta}[p_{\alpha}'(t)] \quad \text{for } all \ t \text{ and } P(t). \tag{40}$$

LEMMA 3. *Suppose that all of the assumptions of the theorem hold. Then, the MGAE reinforcement scheme has the following convergence property under the nonstationary multi-teacher environment (NMT) satisfying condition (5): $p_{\alpha}(t)$ converges with probability 1. Further, let $\lim_{t\to\infty} p_{\alpha}(t) = p_{\infty}^{\alpha}$ with probability 1. Then, $p_{\infty}^{\alpha} = 1$ or 0 with probability 1.*

Proof. $|p_\alpha(t)| \leq 1$ for all t. Then, from Lemma 1, $p_\alpha(t)$ converges with probability 1 (Doob 1953). Now we will prove that $p_\infty^\alpha = 1$ *or* 0 with probability 1. Assume that there is a region D such that $\mu(D) \neq 0$ and $0 < p_\infty^\alpha < 1$ in D. It follows from (30) that

$$E[p_\alpha(t+1)] - E[p_\alpha(t)] \geq \int_\Omega [\lambda+u](1-p_\alpha)p_\alpha[C_\alpha(t)-C_\beta(t)]du \qquad (41)$$

Since $p_\alpha(t)$ converges with probability 1 to p_∞^α and $|p_\alpha(t)| \leq 1$ for all t,

$$\lim_{t\to\infty} E[p_\alpha(t)] = E[p_\infty^\alpha] \qquad (42)$$

Hence,

$$\lim_{t\to\infty} \{E[p_\alpha(t+1)] - E[p_\alpha(t)]\} = \lim_{t\to\infty} E[p_\alpha(t+1)] - \lim_{t\to\infty} E[p_\alpha(t)] = 0 \qquad (43)$$

Let

$$\lambda + u < -G \quad (G>0)$$

Then, from (28),

$$\lim_{t\to\infty} \int_\Omega [\lambda+u](1-p_\alpha)p_\alpha[C_\alpha(t)-C_\beta(t)]du > \lim_{t\to\infty} \int_\Omega \frac{\delta G}{n}(1-p_\alpha)p_\alpha\, du \qquad (44)$$

$$= \int_\Omega \frac{\delta G}{n}(1-p_\infty^\alpha)p_\infty^\alpha du$$

$$= \int_D \frac{\delta G}{n}(1-p_\infty^\alpha)p_\infty^\alpha du > 0$$

It is clear from (41) that (43) is incompatible with (44). Therefore

$$p_\infty^\alpha = 1 \text{ or } 0$$

with probability 1.

Taking advantage of the above three lemmas, the Theorem can easily be proved.

Proof. From Lemma 2,

$$h_{z,\vartheta}[p_\alpha'(0)] \geq \int_\Omega h_{z,\vartheta}[p_\alpha'(1)]\, du \geq \cdots \qquad (45)$$

Consequently,

$$h_{z,\vartheta}[p_\alpha'(0)] \geq \lim_{t\to\infty} \int_\Omega h_{z,\vartheta}[p_\alpha'(t)]\, du \tag{46}$$

Since $|h_{z,\vartheta}[p_\alpha'(t)]|$ is bounded above (≤ 1),

$$\lim_{t\to\infty} \int_\Omega h_{z,\vartheta}[p_\alpha'(t)]\, du = \int_\Omega \lim_{t\to\infty} h_{z,\vartheta}[p_\alpha'(t)]\, du \tag{47}$$

Let

$$p_\infty^\beta = 1 - p_\infty^\alpha \tag{48}$$

Then, from lemma 3,

$$p_\infty^\beta = 0 \text{ } or \text{ } 1 \tag{49}$$

with probability 1. Since $h_{z,\vartheta}(p)$ is a continuous function of p, we obtain the following equality:

$$\lim_{t\to\infty} h_{z,\vartheta}[p_\alpha'(t)] = h_{z,\vartheta}(p_\infty^\beta) \tag{50}$$

with probability 1. Furthermore,

$$0 < h_{z,\vartheta}(p) < 1 \qquad when \ 0<p<1 \tag{51}$$

$$h_{z,\vartheta}(0) = 0, \qquad h_{z,\vartheta}(1) = 1$$

It follows from (49) and (50) that

$$\lim_{t\to\infty} h_{z,\vartheta}[p_\alpha'(t)] = p_\infty^\beta \tag{52}$$

with probability 1. Therefore, from (46), (47), and (52),

$$h_{z,\vartheta}[p_\alpha'(0)] \geq \int_\Omega p_\infty^\beta\, du$$

It is clear that

$$\lim_{\vartheta\to 0} h_{z,\vartheta}[p_\alpha'(0)] = \lim_{\vartheta\to 0} \left[\frac{exp[\frac{z(r-1)}{r\vartheta}]-1}{exp[\frac{z}{\vartheta}]-1} \right] = 0 \tag{54}$$

Hence, from (53) and (54),

$$\lim_{\vartheta\to 0} \lim_{t\to\infty} E[p_\alpha(t)] = 1$$

APPLICATION TO NOISE-CORRUPTED, MULTI-OBJECTIVE PROBLEM

In this section, we consider a parameter self-optimization problem with noise-corrupted, multi-objective functions as an application of learning behaviors of stochastic automata operating in an unknown nonstationary multi-teacher environment.

Suppose the $J_1(\alpha),\ldots,$ and $J_n(\alpha)$ are unknown objective functins of a parameter $\alpha\in[\alpha_1, \ldots, \alpha_r]$ except that they are bounded $(-M\leq J_1(\alpha),\ldots,J_n(\alpha)\leq M)$. It is assumed that measurements of $J_i(\alpha)$ $(i=1,\ldots,n)$ can be obtained only from the noise-corrupted observations.

$$g_i(\alpha,\xi_i) = J_i(\alpha) + \xi_i \quad (i=1,\ldots,n) \tag{55}$$

Here, $J_i(\alpha)$ is assumed to have unique maximum $J_i(\alpha_{\beta_i})$:

$$J_i(\alpha_{\beta_i}) = \max\,[J_i(\alpha_1),\ldots,J_i(\alpha_r)] \tag{56}$$

Each objective function $J_i(\alpha)$ has the claim to be maximized $(i=1,\ldots,n)$. However, generally, the relation $\alpha_{\beta_1} = \alpha_{\beta_2} = \cdots = \alpha_{\beta_n}$ does not hold. This is one of the most difficult points of multi-objective optimization problems.

The learning behaviors of stochastic automata operating in the last section can be used to find an appropriate parameter in this problem. Let us try to identify the ith action y_i of stochastic automaton B with the ith parameter value α_i $(i=1,\ldots,r)$. Choosing the ith parameter α_i at time t corresponds to B producing the output y_i at time t. For simplicity, we consider the stochastic automaton B under P-model environment.

Let k_t^j be a measurement of $g_j(\alpha,\xi_j)$ at time t. Further, let $\bar{k}_t^j(t=0,1,\ldots;\ j=1,\ldots,n)$ be defined as

$$\bar{k}_0^j = \begin{cases} k_0^j & \text{if } -M\leq k_0^j\leq M \\ M & \text{if } k_0^j>M \\ -M & \text{if } k_0^j<-M \end{cases} \tag{57}$$

$$\bar{k}_t^j = \begin{cases} \dfrac{1}{t+1}(t\cdot\bar{k}_{t-1}^j + k_t^j) & \text{if } -M\leq k_t^j\leq M \\ \dfrac{1}{t+1}(t\cdot\bar{k}_{t-1}^j + M) & \text{if } k_t^j>M \\ \dfrac{1}{t+1}(t\cdot\bar{k}_{t-1}^j - M) & \text{if } k_t^j<-M \end{cases}$$

Using these values, we define reward and penalty as follows:

Suppose that $\alpha(t) = \alpha_i$ $(i=1,\ldots,r)$. If $k_l^j \geq \bar{k}_{l-1}^j$, then the stochastic automaton B receives reward response $S_j^i = 0$ from the $j$the teacher R_j $(j=1,\ldots,n)$. (This means that the jth noise-corrupted, objective function $J_j(\alpha)$ gives an affirmative answer to the parameter α_i.)

On the contrary, if $k_l^j < \bar{k}_{l-1}^j$, then the stochastic automaton B receives penalty response $S_j^i = 1$ from the jth teacher R_j $(j=1,\ldots,n)$.

The stochastic automaton chages the state vector $P(t)$ to $P(t+1)$ by the n responses $(S_1^i,\ldots,s_n^i)$ which it has received from the n teachers $R_1, \ldots,$ and R_n.

Now let us consider the learning behavior of B. If the parameter α_i is selected at time t, B receives penalty from the jth teacher R_j with the probability

$$\mu[g_j(\alpha_i,\xi_j(t)) < \bar{k}_{l-1}^j]$$

From (55),

$$\begin{aligned}\mu[g_j(\alpha_i,\xi_j(t)) < \bar{k}_{l-1}^j] &= u[\xi_j(t) < \bar{k}_{l-1}^j - J_j(\alpha_i)] \\ &= p_{\xi_j}[\xi_j(t) < \bar{k}_{l-1}^j - J_j(\alpha_i)]\end{aligned} \tag{58}$$

($p_{\xi_j}(\cdot)$ is the distribution function of ξ_j $(j=1,\ldots,n)$.)

Since $J_j(\alpha)$ is assumed to have unique maximum $J_j(\alpha_{\beta_j})$,

$$u[\xi_j(t) < \bar{k}_{l-1}^j - J_j(\alpha_{\beta_j})] < u[\xi_j(t) < \bar{k}_{l-1}^j - J_j(\alpha)] \tag{59}$$

$$\text{for } \textit{all } \bar{k}_{l-1}^j \text{ and } \textit{all } \alpha\ (\alpha_{\beta_j} \neq \alpha) \quad (j=1,\ldots,n)$$

(See Figure 3.)

Let

$$C_j^i(t,\omega) = \mu[g_j(\alpha_i,\xi_j(t)) < \bar{k}_{l-1}^j] \tag{60}$$

The reason why we use the notation $C_j^i(t,\omega)$ is to represent the probability that stochastic automaton B receives penalty response from the jth teacher when it selects the ith parameter α_i at time t. (Here, $\omega \in \Omega$, Ω being the supporting set of the probability measure space $(\Omega, \tilde{B}, u)$.

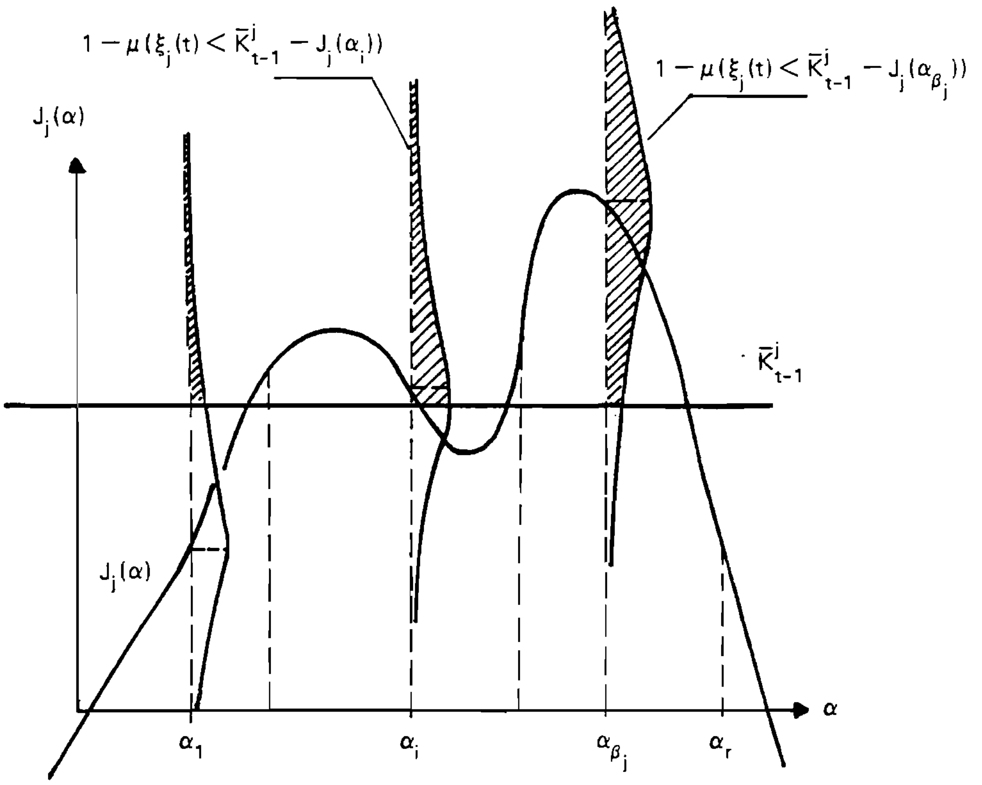
$1 - \mu(\xi_j(t) < \bar{K}^j_{t-1} - J_j(\alpha_i))$
$1 - \mu(\xi_j(t) < \bar{K}^j_{t-1} - J_j(\alpha_{\beta_j}))$
$J_j(\alpha)$
$\bar{K}^j_{t-1}$
$J_j(\alpha)$
α_1
α_i
α_{β_j}
α_r
α

Figure 3. The value of $\{1-\mu[\xi_j(t) < \bar{k}^j_{t-1} - J_j(\alpha_{\beta_j})]\}$.

$$F_t = \sigma[P(0),\ldots,P(t),\bar{k}_0^1,\bar{k}_0^2,\ldots,\bar{k}_0^n,\bar{k}_1^1,\ldots,\bar{k}_1^n,\ldots\bar{k}_t^1,\ldots\bar{k}_t^n]$$

($\sigma[P(0),\ldots,\bar{k}_t^n]$ is the smallest Borel field of ω-sets with respect to which $P(0),\ldots,$ and $\bar{K}_t^n$ are all measurable.) $\tilde{B}$ is the smallest Borel field which contains $\bigcup_{t=0}^{\infty} F_t$. u is the probability measure which satisfies $u(\Omega) = 1$.)

Therefore, it follows from (56), (58), (59), and (60), that

$$C_j^{\beta_j}(t,\omega) + \delta_j < C_j^i(t,\omega) \tag{61}$$

for *all* t, *all* i $(i=1,\ldots,r;\ i\neq\beta_j)$, *all* $\omega \in \Omega$, and *some positive number* δ_j $(j=1,\ldots,n)$

If the strict condition

$$\alpha_{\beta^*} = \alpha_{\beta_1} = \cdots = \alpha_{\beta_n} \tag{62}$$

holds, then it can be easily derived from (61) that

$$C_1^{\beta^*}(t,\omega)+\ldots+C_n^{\beta^*}(t,\omega)+\delta < C_1^i(t,\omega)+\ldots+C_n^i(t,\omega) \tag{63}$$

$$(\delta=\delta_1+\ldots+\delta_n)$$

for *all* t, *all* i $(i=1,\ldots,r;\ i\neq\beta^*)$, *all* $\omega\in\Omega$, and *the positive number* δ

Therefore, using the theoretical results obtained in the last section, we can prove that

$$\lim_{\theta\to 0}\lim_{t\to\infty} E[p_{\beta^*}(t)] = 1$$

is ensured by the MGAE reinforcement scheme.

Even if the strict condition (62) does not hold, the MGAE reinforcement scheme finds the parameter α which satisfies the relation (5). (The result obtained so far is a generalization of the work done by Baba (1978).)

Remark: Although we have used P-model, all of the studies done in the last section can be applied to this case.

CONCLUSION

We have discussed the learning behavior of stochastic automata under the nonstationary multi-teacher environment (NMT) in which penalty strengths are functions of t and ω, where t represents time and ω is a point of the basic ω-space Ω. It has been proved that the MGAE reinforcement scheme, which is an extended form of the GAE reinforcement scheme, ensures ε-optimality under the nonstationary multi-teacher environment (NMT) which satisfies condition (5). We have also considered the parameter self-optimization problem with noise-corrupted, multi-objective functions by stochastic automata and showed that this problem can be reduced to the learning behaviors of stochastic automata operating in the nonstationary multi-teacher environment (NMT) satisfying condition (5).

REFERENCES

Baba, N. and Y. Sawaragi. 1975. On the learning behavior of stochastic automata under a non-stationary random environment. IEEE Trans. Syst., Man, and Cybernetics. 5:273-275.

Baba, N. 1978. Theoretical considerations of the parameter self-optimization by stochastic automata. Int. J. Control. 27:271-276.

Baba, N. 1983. The absolutely expedient nonlinear reinforcement schemes under the unknown multi-teacher environment. IEEE Trans. Syst., Man, and Cybernetics. 13:100-108.

Chandrasekaran, B. and D.W.C. Shen. 1968. On expediency and convergence in variable-structure automata. IEEE Trans. Syst., Sci., and Cybernetics. 4:52-60.

Doob, J.L. 1953. Stochastic Processes. New York: Academic Press.

Koditschek, D.E. and K.S. Narendra. 1977. Fixed structure automata in a multi-teacher environment. IEEE Trans. Syst., Man, and Cybernetics. 7:616-624.

Lakshmivarahan, S. and M.A.L. Thathachar. 1973. Absolutely expedient learning algorithms for stochastic automata. IEEE Trans. Syst., Man, and Cybernetics. 3:281-286.

Narendra, K.S. and M.A.L. Thathachar. 1974. Learning automata--a survey. IEEE Trans. Syst., Man, and Cybernetics. 4:323-334.

Norman, M.F. 1968. On linear models with two absorbing barriers. Journal of Mathematical Psychology. 5:225-241.

Norman, M.F. 1972. Markov Processes and Learning Models. New York: Academic Press.

Sawaragi, Y. and N. Baba. 1973. A note on the learning behavior of variable-structure stochastic automata. IEEE Syst., Man, and Cybernetics. 3:644-647.

Srikantakumar, P.R. and K.S. Narendra. 1982. A learning model for routing in telephone networks. SIAM J. Control and Optimization. 20:34-57.

Thathachar, M.A.L. and B. Bhakthavathsalam. 1978. Learning automaton operating in parallel environments. J. of Cybernetics and Information Science. 1:121-127.

Tsetlin, M.L. 1961. On the behavior of finite automata in random media. Automation and Remote Control. 22:1345-1354.

Varshavskii, V.I. and I.P. Vorantsova. 1963. On the behavior of stochastic automata with variable-structure. Automation and Remote Control. 24:327-333.