A Class of Optimal Performance Directed Probabilistic Automata

LUC P. DEVROYE

***Abstract*—A new type of probabilistic automaton is presented which is optimal, converges quickly, and is flexible in the sense that several characteristics of the unknown random environment can be learned and incorporated in the scheme without affecting the convergence. Emphasis is on the proof of convergence and a theoretical and experimental comparison with other strategy selection procedures.**

I. Problem Formulation

The problem of the sequential selection of the best of N strategies α_i, $i = 1,\cdots,N$, in a stationary and finite random environment is considered. A finite random environment $\mathscr{E}$ is a finite collection of distribution functions on $\mathscr{R}^1$, let us say $\{F_1,\cdots,F_N\}$. The environment is unknown, i.e., no information whatsoever is available concerning the distribution functions F_i, $i = 1,\cdots,N$. At each iteration (epoch) n, one strategy $S_n \in \{\alpha_1,\cdots,\alpha_N\}$ is picked and applied to the unknown random environment which responds with a number Y_n called the loss or the response of the environment.

Given that $S_n = \alpha_i$, Y_n is a random variable distributed at Y^i where Y^i has the distribution function $F_i(\cdot)$ in $\mathscr{R}^1$ and the mean

$$c_i \triangleq \int y \, dF_i(y) = E\{Y^i\} = E\{Y_n \mid S_n = \alpha_i\}, \qquad i = 1,\cdots,N, \tag{1}$$

where c_i is the expected loss (or risk) with strategy α_i. Throughout this paper, it is further assumed that the variances (2) are finite:

$$\begin{aligned}\sigma_i^2 &= \int (y - c_i)^2 \, dF_i(y) = E\{(Y^i - c_i)^2\} \\ &= E\{(Y_n - c_i)^2 \mid S_n = \alpha_i\} \\ &\le \sigma^2 < \infty, \quad i = 1,\cdots,N. \end{aligned} \tag{2}$$

Manuscript received February 26, 1974; revised July 1, 1975, and June 17, 1976. This work was supported in part by the Air Force under Grant AFOSR 72-2371.

The author is with the Department of Electrical Engineering, University of Texas, Austin, TX 78712.

To describe the strategy selection process, we assume that there exists a probability distribution on $\{\alpha_1,\cdots,\alpha_N\}$, let us say $P_n = (P_{1,n},\cdots,P_{N,n})$ so that

$$\sum_{i=1}^{N} P_{i,n} = 1, \qquad n = 1,2,\cdots \tag{3}$$

$$P_{i,n} = P\{S_n = \alpha_i\}, \qquad i = 1,\cdots,N, \quad n = 1,2,\cdots. \tag{4}$$

The selection probability vector P_n is usually a random vector. The *a priori* expected loss at epoch n is denoted by M_n and is defined by (5):

$$M_n = E\{Y_n \mid P_n\} = \sum_{i=1}^{N} E\{Y^i\}P\{S_n = \alpha_i \mid P_n\} = \sum_{i=1}^{N} P_{i,n}c_i. \tag{5}$$

A set of rules for computing P_{n+1} given (P_j,S_j,Y_j), $j = 1,\cdots,n$, is called a probabilistic automaton [2], [3], [5]–[16], [19]–[21]. In particular, we are interested in optimal automata, i.e., automata which insure that

$$\lim_n E\{M_n\} = d = \min\{c_1,\cdots,c_N\}. \tag{6}$$

For general environments, none of the known probabilistic (stochastic) automata with a variable structure (SAVS) is optimal. To describe the properties of some of the SAVS, Viswanathan and Narendra [11]–[12] introduced the concept of ε-optimality. The underlying idea is the following: For all $\varepsilon > 0$, a set of rules for computing P_{n+1} can be found within the class of stochastic automata under consideration such that $\lim_n \sup E\{M_n\} < d + \varepsilon$. For most of the classes of automata, the set of rules is completely determined by the choice of one parameter.

If $Y^i \in \{0,1\}$ for all i, ε-optimality was proved for a broad class of SAVS [3], [11], [12], [19]–[20]. The most promising experimental results for general environments ($Y^i \in R^1$ for all $i = 1,\cdots,N$) were obtained by Shapiro and Narendra [3], but their procedure is not optimal. The stochastic approximation type automaton of Fu and Nikolic [10] is proved to be optimal, but one of the conditions of convergence is not *a priori* controllable. Experimental results show that their algorithm behaves very similarly to ε-optimal automata.

We will present a new class of probabilistic automata that is optimal, converges quickly, and is flexible in the sense that several characteristics of the unknown environment can be learned. This information can be incorporated in the scheme without affecting the convergence.

II. A New Class of Algorithms

Let $S_n \in \{\alpha_1,\cdots,\alpha_N\}$ be the strategy selected at epoch n and Y_n be the corresponding observed loss. We will need the following random variables which are defined for all $i = 1,\cdots,N$:

$$R_{i,n} = I\{S_n = \alpha_i\} \qquad L_{i,n} = \sum_{k=1}^{n} R_{i,k}$$

$$G_{i,n} = \frac{1}{L_{i,n}} \sum_{k=1}^{n} Y_k R_{i,k} \qquad D_{i,n} = \frac{1}{L_{i,n}} \sum_{k=1}^{n} Y_k^2 R_{i,k}$$

$$V_{i,n} = D_{i,n} - G_{i,n}^2 \tag{7}$$

where $I\{\cdot\}$ is the indicator function of the event $\{\cdot\}$.

All these quantities can be computed recursively. $L_{i,n}$ counts the number of times strategy α_i was applied to the environment up to time n. $G_{i,n}$ is the estimate at epoch n of $c_i = E\{Y^i\}$. $D_{i,n}$ estimates $E\{(Y^i)^2\}$, and $V_{i,n}$ is the nth epoch estimate of σ_i^2. Denote the N-dimensional column vectors corresponding to these quantities by R_n, L_n, G_n, D_n, and V_n. Further, α_{H_n} is the best estimate of the optimal strategy at epoch n and is defined as follows. Let Γ_n be the set of integers j for which $G_{j,n} = \min_i \{G_{i,n}\}$. Choose H_n at random from Γ_n. Clearly,

$$G_{H_n,n} = \min_i \{G_{i,n}\} \qquad H_n \in \Gamma_n \subseteq \{1,\cdots,N\}. \tag{8}$$

To simplify our notation, let J_n denote the N-dimensional column vector with components $J_{i,n} = I\{H_n = i\}$. The random variables in whose asymptotical behavior we are interested are M_n (5), which is the loss to be expected at epoch $n + 1$, and $Z_n = c_{H_n}$, which is the risk associated with α_{H_n}.

A probability vector is, in our context, an N-dimensional column vector from $[0,1]^N$ whose components sum to 1. Let $B_n = (B_{1,n},\cdots,B_{N,n})^T$ be an arbitrary probability vector, $a \in (0,1]$, and let $\{\beta_n\}_{n=1}^{\infty}$ be a random sequence from $[0,1]$ satisfying $\sum_{n=1}^{\infty} \beta_n = \infty$ with probability one and $\lim_n \beta_n = 0$ with probability 1. Then relation (9) defines a PDPA (performance directed probabilistic automaton):

$$P_{i,n+1} = I\{i = n + 1\}, \qquad 0 \le n \le N - 1$$

$$P_{n+1} = \beta_n \left(\frac{a}{N} \mathbf{1} + (1 - a)B_n\right) + (1 - \beta_n)J_n, \qquad N \le n \tag{9}$$

where $\mathbf{1} = (1,1,\cdots,1)^T$.

The reader can easily verify that J_n, P_n, and $1/N \cdot \mathbf{1}$ are probability vectors. B_n is to be picked in such a way that the algorithm performs "well" (e.g., displays a high rate of convergence or samples all the strategies so as to equalize the variances of the estimates $G_{i,n}$, $i = 1,\cdots,N$, etc.). We now give a few examples, some of which gave excellent experimental results. Let $\bar{G}_n = \max_i \{G_{i,n}\}$, $\theta > 0$ and $b > 0$, and let $\{\theta_n\}_{n=1}^{\infty}$ be a sequence from $[0,\infty]$. Define for $i = 1,\cdots,N$ the random variables $C_{i,n}$:

$$\begin{aligned}
&\text{i)} \quad C_{i,n} = 1\\
&\text{ii)} \quad C_{i,n} = (V_{i,n} \cdot L_{i,n}^{-1})^{\theta}\\
&\text{iii)} \quad C_{i,n} = (\bar{G}_n - G_{i,n})^{\theta}\\
&\text{iv)} \quad C_{i,n} = 1 + b(1 - G_{i,n})L_{i,n}\\
&\text{v)} \quad C_{i,n} = (b + L_{i,n}V_{i,n}^{-1}(G_{i,n} - G_{H_n,n})^2)^{-\theta_n}.
\end{aligned} \tag{10}$$

Further, in each case let

$$B_{i,n} = C_{i,n} \left[\sum_{j=1}^{N} C_{j,n}\right]^{-1}$$

so that B_n is a probability vector. If i) is chosen, information is voluntarily rejected and uniform sampling is favored. Therefore, i) is only useful in high equal noise problems, i.e., all σ_i^2 are approximately equal and much larger than $\max_{i,j} |c_i - c_j|$. Choice ii) results in a higher sampling rate for relatively less sampled strategies (i.e., strategies with high estimated variance $V_{i,n}$ relative to the sample size $L_{i,n}$). Choice iii) favors "promising" (low $G_{i,n}$) strategies over other strategies. The parameter θ controls the value of the largest $B_{i,n}$. Choice iv) is very similar to choice iii) and was first suggested by Meerkov [13]. It is only applicable if $Y^i \in [0,1]$ for all i and will be discussed later. The last choice combines the features of ii) and iii) and made algorithm (9) converge quickly in most of the test examples (for particular choices of $\{\theta_n\}_{n=1}^{\infty}$). It is clear that this list (10) is not exhaustive and that, regarding the selection of B_n, the designer is only limited by his imagination and experience.

Besides this class of PDPA, another SAVS will also be dealt with in this paper. Let $\{\zeta_n\}_{n=0}^{\infty}$ and $\{\delta_n\}_{n=0}^{\infty}$ be random sequences from $[0,1]$ satisfying (11):

$$\begin{aligned} &\sum_{n=0}^{\infty} \delta_n = \infty, && \text{with probability 1} \\ &\lim_n \zeta_n = 1, && \text{with probability 1} \\ &\lim_n \delta_n = 0, && \text{with probability 1.} \end{aligned} \tag{11}$$

Then we define this SAVS by

$$P_0 = \frac{1}{N} \cdot 1$$

$$P_{n+1} = P_n + \zeta_n \cdot \left[(1 - \delta_n)J_n + \frac{\delta_n}{N} \cdot 1 - P_n\right], \qquad n \geq 0. \tag{12}$$

Algorithm (12) coincides in form with the algorithm of Fu and Nikolic [10] for $\delta_n = 0$.

III. Proof of Convergence

Both classes of algorithms satisfy the following requirements. There exist random sequences $\beta_{1,n}, \cdots, \beta_{N,n}$ and γ_n, $n = 1,2,\cdots$ with $\lim_n \gamma_n = 1$ with probability 1,

$$\sum_{n=1}^{\infty} \beta_{i,n} = \infty$$

with probability 1 for all $i = 1,\cdots,N$ and

$$\begin{aligned} &P_{i,n} \geq \beta_{i,n}, && i = 1,\cdots,N, \quad n = 1,2,\cdots \\ &P_{H_n,n} \geq \gamma_n, && n = 1,2,\cdots. \end{aligned} \tag{13}$$

For all automata which satisfy the foregoing condition in environments satisfying requirements (1) and (2), (14)–(17) hold:

$$Z_n \to d, \qquad \text{with probability 1 as } n \to \infty \tag{14}$$

$$E\{Z_n\} \to d, \qquad \text{as } n \to \infty \tag{15}$$

$$M_n \to d, \qquad \text{with probability 1 as } n \to \infty \tag{16}$$

$$E\{M_n\} \to d, \qquad \text{as } n \to \infty \tag{17}$$

Remark: Equation (13) implies that $\lim_{n\to\infty} \beta_{i,n} = 0$ with probability 1 for $i = 1,2,\cdots,N$.

Proof: Let C_1 denote the event $[\lim_n \gamma_n = 1]$ and

$$C_2 = \bigcap_{i=1}^{N} \left[\sum_{n=1}^{\infty} \beta_{i,n} = \infty\right].$$

We know that $P\{C_1 \cap C_2\} = 1$. Assume first that $Z_n \to d$ with probability 1 as $n \to \infty$. Because $d \leq c_i \leq M < \infty$ for all $i = 1,\cdots,N$, $E\{Z_n\} \to d$ follows. Next, pick $\varepsilon > 0$ and $\delta > 0$, arbitrarily. We have

$$\begin{aligned} P&\left\{\bigcup_{k\geq n} \left[\sum_{i=1}^{N} P_{i,k}c_i > d + \varepsilon\right]\right\} \\ &\leq P\left\{\bigcup_{k\geq n} [\gamma_k \cdot Z_k + (1 - \gamma_k) \cdot M > d + \varepsilon]\right\} \\ &\leq P\left\{\bigcup_{k\geq n} \left[Z_k > d + \frac{\varepsilon}{2}\right]\right\} + P\left\{\bigcup_{k\geq n} \left[1 - \gamma_k > \frac{\varepsilon}{2M}\right]\right\} \\ &\leq P\left\{\bigcup_{k\geq n} \left[Z_k > d + \frac{\varepsilon}{2}\right]\right\} + \frac{\delta}{2} \leq \delta \end{aligned}$$

which holds for all n large enough in view of $P\{C_1\} = 1$ and $Z_n \to d$ with probability 1. Thus $M_n \to d$ with probability 1, and because $d \leq M_n \leq M < \infty$, $E\{M_n\} \to d$ as well. Therefore, (14) implies (15)–(17), and we will now prove that $Z_n \to d$ with probability 1. Again, pick $\varepsilon > 0$ and $\delta > 0$ arbitrarily. Then

$$\begin{aligned} P&\left\{\bigcup_{k\geq n} [Z_k > d + \varepsilon]\right\} \\ &\leq P\left\{\bigcup_{k\geq n} \bigcup_{i=1}^{N} \left[|G_{i,k} - c_i| > \frac{\varepsilon}{2}\right]\right\} \\ &\leq \sum_{i=1}^{N} P\left\{\bigcup_{k\geq n} |G_{i,k} - c_i| > \frac{\varepsilon}{2}\right\} \\ &\leq \sum_{i=1}^{N} P\left\{\bigcup_{k\geq n} \left[|G_{i,k} - c_i| > \frac{\varepsilon}{2}\right], L_{i,n} \geq R_0\right\} \\ &\quad + \sum_{i=1}^{N} P\{L_{i,n} < R_0\} \\ &\leq \sum_{i=1}^{N} \frac{2\sigma_i^2}{(\varepsilon/2)^2 \cdot R_0} + \sum_{i=1}^{N} P\{L_{i,n} < R_0\} \\ &< \delta/2 + \sum_{i=1}^{N} P\{L_{i,n} < R_0\} \end{aligned} \tag{18}$$

where we used the Hajek–Renyi inequality [24] (in view of the independence of $Y_1,Y_2,\cdots$) and where the constant R_0 is chosen such that

$$R_0 > \frac{16}{\delta\varepsilon^2} \cdot \sum_{i=1}^{N} \sigma_i^2. \tag{19}$$

From the sampling procedure we see that there exist random variables $L_{i,n}^*$ with

$$L_{i,n} \geq L_{i,n}^* = \sum_{j=1}^{n} Z_{i,j}$$

where $Z_{i,1},\cdots,Z_{i,n}$ are independent binary valued random variables with $P\{Z_{i,j} = 1\} = \beta_{i,j}$. Notice that

$$E\{L_{i,n}^*\} = \sum_{j=1}^{n} \beta_{i,j}$$

and

$$\sigma^2\{L_{i,n}^*\} \leq \sum_{j=1}^{n} \beta_{i,j}.$$

For all n sufficiently large, i.e., such that $E\{L_{i,n}^*\}$ is greater than $2R_0$ and greater than $8N/\delta$, we obtain by means of Chebyshev's inequality:

$$\begin{aligned} P\{L_{i,n} < R_0\} &\leq P\{L_{i,n}^* - E\{L_{i,n}^*\} < R_0 - E\{L_{i,n}^*\}\} \\ &\leq P\{L_{i,n}^* - E\{L_{i,n}^*\} < -E\{L_{i,n}^*\}/2\} \\ &\leq \frac{\sigma^2\{L_{i,n}^*\}}{(E\{L_{i,n}^*/2\})^2} \\ &\leq 4\left(\sum_{j=1}^{n} \beta_{i,j}\right)^{-1} < \delta/2N. \end{aligned} \tag{20}$$

Collecting bounds gives that for all n sufficiently large,

$$P\left\{\bigcup_{k\geq n} [Z_k > d + \varepsilon]\right\} < \delta,$$

which completes the proof.

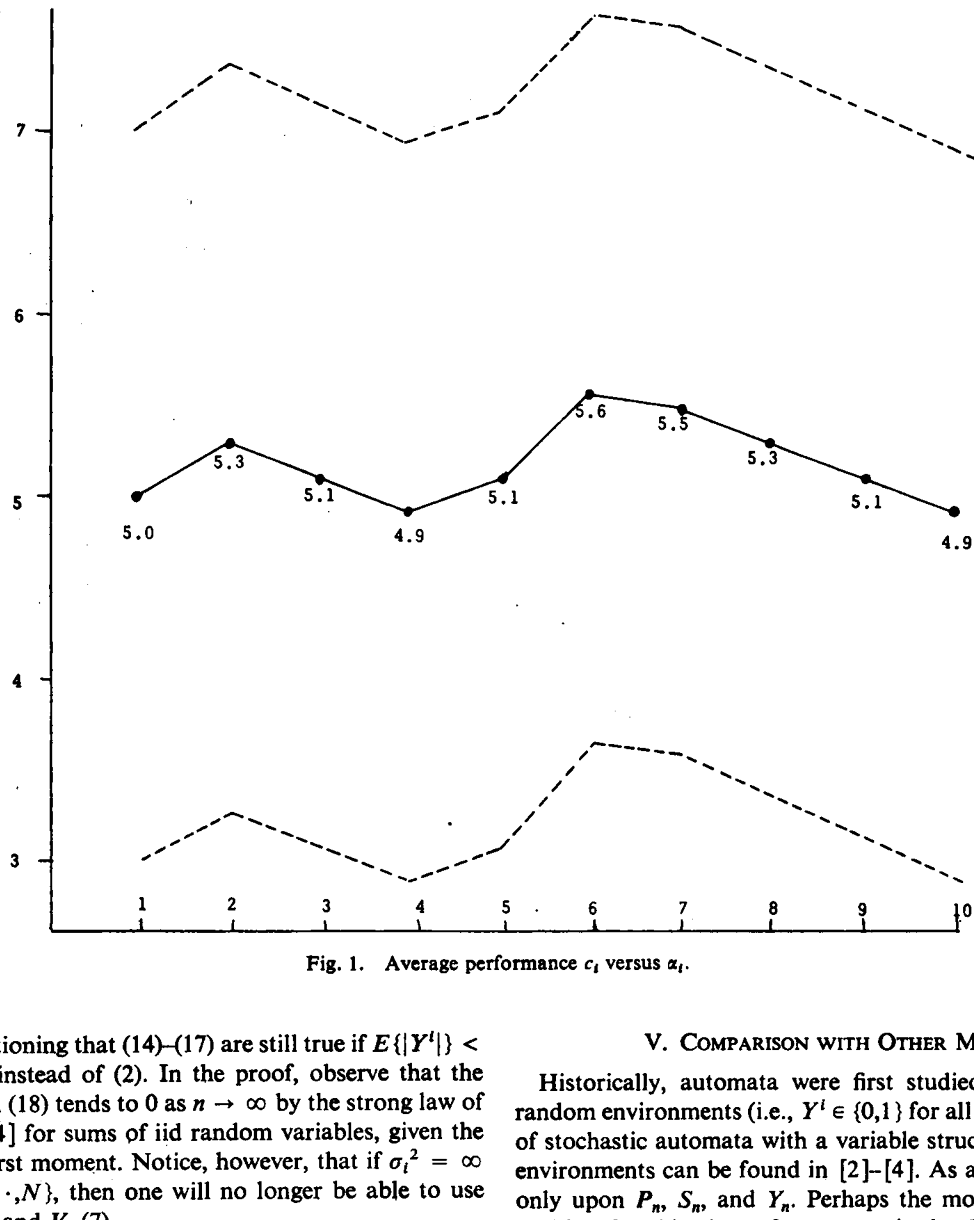

Fig. 1. Average performance c_i versus α_i.

It is worth mentioning that (14)–(17) are still true if $E\{|Y^i|\} < \infty$, $i = 1,\cdots,N$, instead of (2). In the proof, observe that the first probability in (18) tends to 0 as $n \to \infty$ by the strong law of large numbers [24] for sums of iid random variables, given the existence of the first moment. Notice, however, that if $\sigma_i^2 = \infty$ for some $i \in \{1,\cdots,N\}$, then one will no longer be able to use the estimators D_n and V_n (7).

IV. Epsilon-Optimality

With the following fixed-parameter algorithm (21),

$$P_0 = N^{-1} \cdot \mathbf{1}$$

$$P_{n+1} = \beta_0 N^{-1} \cdot \mathbf{1} + (1 - \beta_0) J_n, \qquad \text{for } n \geq 0 \tag{21}$$

where $\beta_0 \in (0,1]$, there is at every instant a positive probability of selecting each strategy. Hence this scheme (or similar more complicated ones) can be used in nonstationary environments, provided the performance is not evaluated by simple averaging but by a finite memory device such as an exponential filter. From the previous theorem we know that since $P\{C_2\} = 1$, $\lim_n Z_n = d$ with probability one. Thus

$$\lim_n E\{M_n\} = d + \beta_0 \cdot N^{-1} \cdot \sum_{i=1}^{N} (c_i - d).$$

To achieve ε-optimality, it suffices to take

$$\beta_0 < \varepsilon / N^{-1} \sum_{i=1}^{N} (c_i - d).$$

V. Comparison with Other Methods

Historically, automata were first studied for use in $0 - 1$ random environments (i.e., $Y^i \in \{0,1\}$ for all i). Excellent surveys of stochastic automata with a variable structure for use in such environments can be found in [2]–[4]. As a rule, P_{n+1} depends only upon P_n, S_n, and Y_n. Perhaps the most representative algorithm for this class of automata is the L_{R-I} (linear reward-inaction) scheme (22):

$$P_0 = N^{-1} \cdot \mathbf{1};\ P_{n+1} = P_n + \beta_0(1 - Y_n)(R_n - P_n),$$
$$\text{for } n \geq 0, \quad \beta_0 \in (0,1). \tag{22}$$

This algorithm is ε-optimal [11], [19], [20], and its main advantage is that it requires no performance estimation through some auxiliary variables as G_n. The deceleration algorithm of Meerkov [13] is a completely different type of algorithm for $0 - 1$ environments. In fact, his algorithm is exactly of the form (9) with choice (10, iv) for B_n, where $a = 0$ and $\beta_n = 1$ for all n. However, he proves only convergence in probability, and experiments show that for any choice of the parameter b, his procedure converges extremely slowly. If $Y^i \in [0,1]$ for all i (or if $Y^i \in [A,B]$ for all i and A,B are known, we can consider $Y^i - A/B - A$), (22) can still be used, but convergence or even ε-optimality are not yet established (see [7], [14]). If A and B are unknown, Viswanathan proposes estimation of A and B by A_n and B_n as the search proceeds [7] and the use of $Y_n^* = Y_n - A_n/B_n - A_n$ in (22). As an example of a SAVS which is not even ε-optimal but merely "expedient" [16], [21] (for the

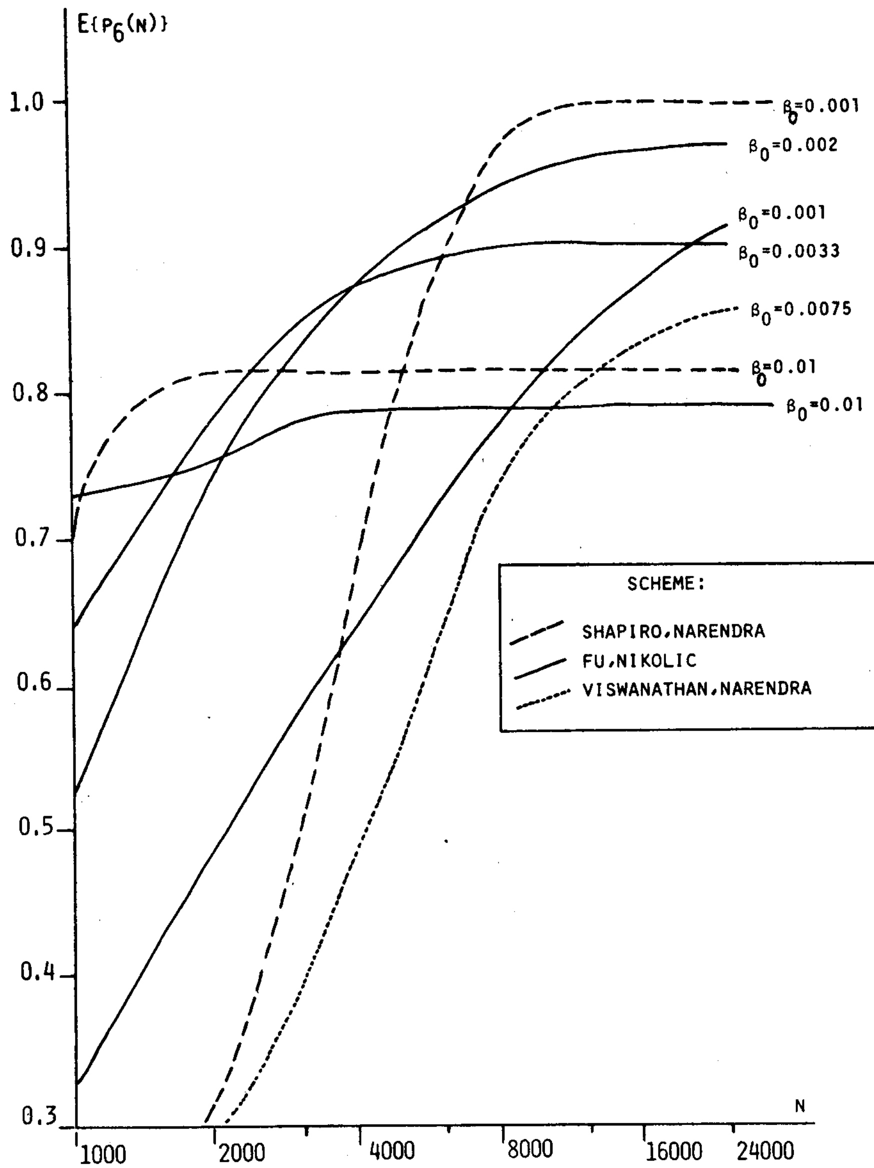

Fig. 2. $E\{P_6(N)\}$ versus N for stochastic automata with variable structure.

latest definition, see [5]), we cite the $L_{R\text{-}P}$ (linear-reward-penalty) automaton of Fu and Maclaren [4], [8], [15]:

$$P_0 = N^{-1} \cdot \mathbf{1};$$

$$P_{n+1} = P_n + \beta_0 \left[R_n(1 - Y_n) + (1 - R_n) \frac{Y_n}{N-1} - P_n \right],$$

$$\text{for } n \geq 0, \quad \beta_0 \in (0,1). \quad (23)$$

It is shown in [4], [8], [15] that for [0,1] environments,

$$\lim_n E\{P_{k,n}\} = c_k^{-1} \left(\sum_{i=1}^{N} c_i^{-1} \right)^{-1}$$

so that asymptotic optimality is, in general, excluded.

If $Y^i \in (-\infty, +\infty)$ for all i, one can try to find some nonlinear transformation, let us say g: $(-\infty, +\infty) \rightarrow [0,1]$ or h: $(-\infty, +\infty) \rightarrow \{0,1\}$ and use $g(Y_n)$ or $h(Y_n)$ in the schemes for $0 - 1$ or [0,1] environments. In general, however, such transformations do not preserve the order in the c_i; i.e., if $c_i = E\{Y^i\} < E\{Y^j\} = c_j$, then it is desired that $c_i^* = E\{g(Y^i)\} < E\{g(Y^j)\} = c_j^*$. It should be pointed out that for most of the environments an order preserving transformation $h(Y)$ exists. For instance, let Y^i be Gaussian with mean c_i and nonzero variance σ^2, then for any $\lambda \in R^1$: $h(Y) = I\{Y > \lambda\}$ is order preserving. If no assumptions can be made about the environment, one really needs a "memory" for past measurements (such as G_n, etc.). The solutions presented by various authors all had the same characteristic, i.e., P_{n+1} depends upon P_n and J_n. Further, ε-optimality is the best that can be said about the asymptotical behavior of some of the SAVS. Moreover, due to the fixed parameter β_0 (see (22) and (23)), there exists a positive probability that $M_n \nrightarrow d$ as $n \rightarrow \infty$ (because of "absorption" in one of the states $\alpha_1, \cdots, \alpha_N$ of the automaton), although this probability can be made arbitrarily small by decreasing β_0. In general environments, very high rates of convergence can be observed with the $L_{R\text{-}I}$ type algorithm of Shapiro and Narendra [3]:

$$P_0 = N^{-1} \cdot \mathbf{1}; \, P_{n+1} = P_n + \beta_0(R_n^T J_n)(J_n - P_n),$$

$$\text{for } n \geq 0 \text{ and } \beta_0 \in (0,1). \quad (24)$$

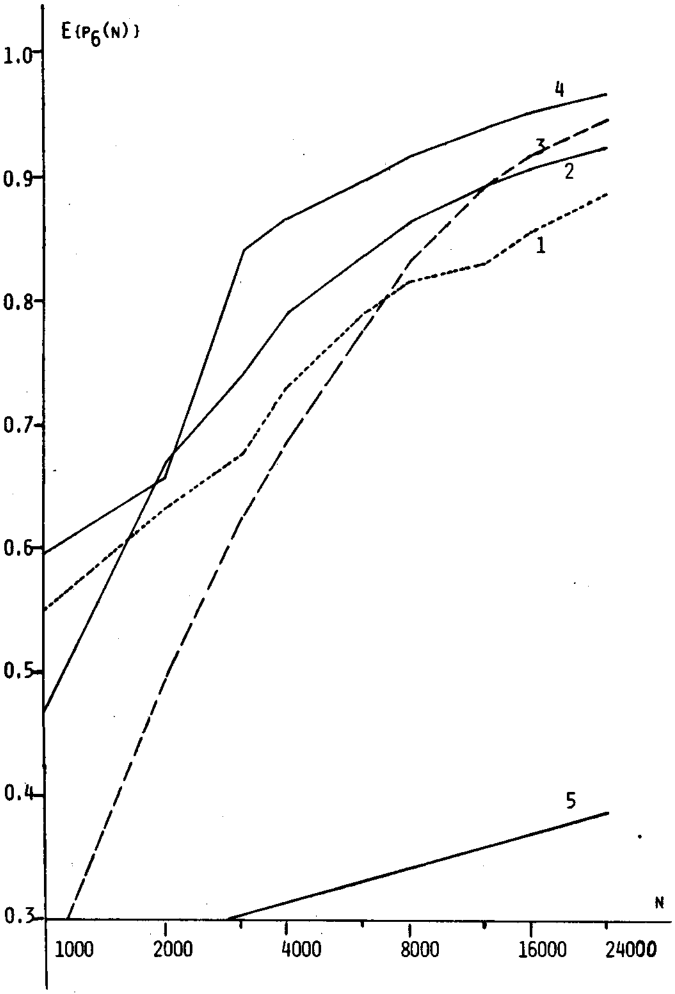

Fig. 3. $E\{P_6(N)\}$ versus N for performance-directed probabilistic automata.

The idea is to update P_n only if the selected strategy S_n was the right choice, i.e., $S_n = \alpha_{H_n}$. The scheme behaves experimentally in an ε-optimal way but is not known to be ε-optimal. In any case, (24) is not optimal in the sense of (6). Fu and Nikolic [10] proposed a stochastic approximation type of algorithm with varying parameter:

$$P_0 = N^{-1} \cdot \mathbf{1};\ P_{n+1} = P_n + \beta_n(J_n - P_n),$$
$$\text{for } n \geq 0 \text{ and } \{\beta_n\}_{n=0}^{\infty} \subset [0,1]. \quad (25)$$

They prove that if $\sigma^2 < \infty$ and $c_1 < c_2 < \cdots < c_N$, then $P_{1,n} \to 1$ with probability 1, provided the sequence $\{\beta_n\}_{n=0}^{\infty}$ satisfies:

$$\sum_{n=1}^{\infty} \beta_n = \infty \qquad \sum_{n=1}^{\infty} \beta_n^2 < \infty$$

and

$$\sum_{n=1}^{\infty} \beta_n P\{H_n = i \mid Y_1, \cdots, Y_{n-1}\} < \infty, \qquad \text{for all } i = 2, \cdots, N.$$

Unfortunately, this latter condition is difficult to check in advance. In fact, it is not hard to see that it is a very strict condition. Experiments with several environments have shown that the algorithm behaves ε-optimally, much like the L_{R-I} algorithm (24).

The main differences between the PDPA and the discussed algorithms (22)–(25) are the following.

1) The bulk of the information is not carried by P_n but by G_n. Note that the same is true for algorithm (12) since $\zeta_n \to 1$ with probability 1 so that asymptotically, $P_{n+1} \approx (1 - \delta_n)J_n + \delta_n N^{-1} \cdot \mathbf{1}$. As a consequence, the convergence properties of G_n can be exploited to prove the with probability 1 convergence of the algorithms (9) and (12).

2) The second difference is the freedom in the design of a technique (through B_n, see Section II).

VI. Experiments

Consider the maximization problem of Narendra [3] with $N = 10$. Fig. 1 gives the plot of c_i versus α_i. Y^i is uniform on $[c_i - 2, c_i + 2]$. Note that α_6 is the optimal strategy. Obviously, the problem can be regarded as a minimization problem if Y_n is replaced by $-Y_n$. Figs. 2 and 3 depict multiple run estimates of $E\{P_{6,n}\}$ versus n for various techniques.

In Fig. 2, the results are given for three SAVS:

1) the algorithm of Shapiro and Narendra (24) with $\beta_0 = 0.01$ and 0.001 (the curves are averages of 47 runs),
2) the stochastic approximation type algorithm of Fu and Nikolic (25) with $\beta_n = \beta_0(1 + 0.001n)^{-1}$ and $\beta_0 = 0.01$, 0.0033, 0.002, and 0.001 (the curves are averages of 43 runs),
3) the L_{R-I} type automaton with adaptive A_n and B_n of Viswanathan and Narendra (β_0 was 0.0075). (See [7] for a detailed description).

For all these algorithms there exists a nonoptimal asymptotic level for $E\{P_{6,n}\}$, and this level can be brought arbitrarily close to 1 by decreasing the value of β_0 at the expense of a slower rate of convergence. Note also that the SAVS (23) of Maclaren and Fu is not competitive at all in such high-noise situations; e.g., if Y_n is replaced in (23) by $(Y_n - 2.7)/(7.6 - 2.7)$, $\lim E\{P_{6,n}\} = 0.121$ follows.

The results with some of the PDPA are given in Fig. 3. They show, in general, less sensitivity with respect to the choice of β_0 and behave asymptotically optimally. Curves 1, 2, and 3 are 40-run averages for algorithm (9) with

$$\beta_n = \min\left\{1; \frac{\beta_0}{(1 + 0.001n)}\right\}$$

and $\beta_0 = 0.5$, 1.0, and 1.5, respectively. B_n is constructed with the aid of (10, i). Note that the gradient of the curves even for $n = 24\,000$ is still very steep. Curve 4 is a ten-run average for algorithm (9) with the same β_n and $\beta_0 = 0.1$. B_n is constructed with the aid of the more complex choice (10, v) for $C_{i,n}$, $i = 1, \cdots, N$. We took $a = 0.005$, $b = 1.0$, $\theta_n = \beta_n^{-1}$. The initial rate of convergence for the PDPA is slower than the rate for some SAVS, but this phenomenon almost inevitably occurs when asymptotically optimal systems are compared with nonoptimal systems. Note further that the main field of application of the PDPA (9) (10, v) is where the on-line estimation of σ_i^2, $i = 1, \cdots, N$, can pay off, i.e., in much more irregular noise situations than the one of the test example. Curve 5 in Fig. 3 is a 50-run average with the deceleration algorithm of Meerkov (i.e., algorithm (9), (10, iv) with $a = 0$, $\beta_n = 1$ for all n). For all $b > 5$, $E\{P_{6,n}\}$ was below 0.3 even after 24 000 units of time. For small b, the behavior was as in Fig. 3, curve 5 (where $b = 1$). Y_n was replaced by $Y_n^* = I\{Y_n > 5.18\}$ which, in this example is an order-preserving transformation.

VII. Conclusion

A new type of automata is developed for general environments ($Y^i \in (-\infty, +\infty)$ for all i), differing from stochastic automata with a variable structure in that the selection probabilities are not adaptive but depend directly on some estimated parameters. One of the advantages of the PDPA is that with probability 1 convergence is insured for all the random variables of special interest. Furthermore, the PDPA is remarkably flexible regarding learning some parameters of the environment and introducing this learned information in the scheme. Examples are given of mean and variance estimators as well.

The asymptotic optimality of the scheme was proved and can be extended without too much effort towards strategy selection problems with a countably infinite number of strategies. In addition, it is indicated how the PDPA can be modified for use in nonstationary environments. Finally, some PDPA are compared both theoretically and experimentally with most of the well-known SAVS.

References

[1] H. Robbins, M. Sobel, and N. Starr, "A sequential procedure for selecting the largest of k means," *Ann. Math. Stat.*, vol. 39, no. 1, pp. 88–92, 1968.

[2] Ya. Z. Tsypkin and A. S. Poznyak, "Finite learning automata," *Engineering Cybernetics*, vol. 10, no. 3, pp. 479–490, 1972.

[3] I. J. Shapiro and K. S. Narendra, "Use of stochastic automata for parameter self-optimization with multimodal performance criteria," *IEEE Trans. Syst. Sci. and Cybernetics*, vol. SSC-5, no. 4, pp. 352–360, 1969.

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

[5] S. Lakshmivarahan and M. A. L. Thathachar, "Absolutely expedient learning algorithms for stochastic automata," *IEEE Trans. Syst., Man, and Cybernetics*, vol. SMC-3, no. 3, pp. 281–286, 1973.

[6] I. P. Vorontsova, "Algorithms for changing stochastic automata transition probabilities," *Probl. Peredach. Inform.*, vol. 1, no. 3, pp. 122–126, 1965.

[7] R. Viswanathan and K. S. Narendra, "Stochastic automata models with applications to learning systems," *IEEE Trans. Syst., Man, and Cybernetics*, vol. SMC-3, no. 1, pp. 107–111, 1973.

[8] K. S. Fu and R. W. Maclaren, "An application of stochastic automata to the synthesis of learning systems," Purdue Univ., Lafayette, Indiana, Tech. Rep., TR-EE65-17, 1965.

[9] G. J. McMurtry and K. S. Fu, "A variable structure automaton used as a multimodal searching technique," *IEEE Trans. Automat. Control*, vol. AC-11, no. 3, pp. 379–387, 1966.

[10] K. S. Fu and Z. J. Nikolic, "On some reinforcement techniques and their relation to the stochastic approximation," *IEEE Trans. Automat. Control*, vol. AC-11, no. 4, pp. 756–758, 1966.

[11] R. Viswanathan and K. S. Narendra, "A note on the linear reinforcement scheme for variable-structure stochastic automata," *IEEE Trans. Syst., Man, and Cybernetics*, vol. SMC-2, no. 2, pp. 292–294, 1972.

[12] ——, "Application of stochastic automata models to learning systems with multimodal performance criteria," Becton Center, Yale Univ., New Haven, Conn., Tech. Rep. CT-40, 1972.

[13] S. M. Meerkov, "Deceleration in the search for the global extremum of a function," *Automation and Remote Control*, vol. 33 no. 12, pp. 2029–2037, 1972.

[14] L. G. Mason, "An optimal learning algorithm for *S*-model environments," *IEEE Trans. Automat. Control*, vol. AC-18, no 5, pp. 493–496, 1973.

[15] R. W. Maclaren, "A stochastic automaton model for the synthesis of learning systems," *IEEE Trans. Syst., Sci. and Cybernetics*, vol. SSC-2, no. 4, pp. 109–114, 1966.

[16] B. Chandrasekaran and D. W. C. Shen, "On expediency and convergence in variable-structure automata," *IEEE Trans. Syst., Sci. and Cybernetics*, vol. SSC-4, no. 1, pp. 52–60, 1968.

[17] M. Fisz, *Probability Theory and Mathematical Statistics*. New York: John Wiley and Sons, 1963.

[18] L. D. Cockrell and K. S. Fu, "On search techniques in adaptive systems," Purdue Univ., Lafayette, Indiana, Tech. Rep. TR-EE70-01, pp. 55–56, 1970.

[19] Y. Sawaragi and N. Baba, "A note on the learning behavior of variable-structure stochastic automata," *IEEE Trans. Syst., Man, and Cybernetics*, vol. SMC-3, no. 6, pp. 644–647, 1973.

[20] ——, "The ε-optimal nonlinear reinforcement schemes for stochastic automata," *IEEE Trans. Syst., Man, and Cybernetics*, vol. SMC-4, no. 1, pp. 126–131, 1974.

[21] M. Tsetlin, "On the behavior of finite automata in random media," *Automation and Remote Control*, vol. 22, no. 10, pp. 1345–1354, 1961.

[22] J. Lamperti, *Probability*. W. A. Benjamin, Inc., pp. 41–42, 1966.

[23] A. S. Poznyak, "Learning automata in stochastic programming problems," *Automation and Remote Control*, vol. 34, no. 9, part 2, pp. 1608–1619, 1973.

[24] B. V. Gnedenko, *The Theory of Probability*. New York: Chelsea Publishing Co., 1967.