Probabilistic Search as a Strategy Selection Procedure

LUC P. DEVROYE

Abstract—An alternative solution to the problem of the selection of the best strategy in a random environment is presented by using a probabilistic search procedure. The asymptotic optimality of the technique is proved, and a brief comparison with stochastic automata with variable structures is made. A specific organization of the optimal search procedure is developed based on continued learning of some statistics of the random environment, and it is shown to be fast-converging, powerful in high noise random environments, and insensitive to search parameter selection.

I. Introduction

The problem of the selection of the best strategy in a random environment has been extensively dealt with by using stochastic automata with variable structures (SAVS) [1]–[12]. The SAVS approach, efficient in P-model environments [2], [5], [7], [8], [10], [12] has recently been used also for S-model [3], [9], [11] and general environments [6]. The inconvenience with SAVS is that although the selection probabilities in the automata converge with probability one to zero or one [8], [12], [13], they do not always converge to the desired value. Therefore, the concept of ε-optimality had to be introduced [8], [12] which is weaker than convergence in probability. Many experiments have shown that the SAVS looses its attractiveness when the number of strategies M is very large and the noise on the output ("response") of the random environment is large. In SAVS, all the information concerning past measurements is stored in a set of probabilities, and valuable data are wasted. One can expect that by enlarging the memory and processing more data as they come in, an acceleration of the rate of convergence can be obtained.

Manuscript received July 1, 1975; revised October 27, 1975. This work was supported in part by Air Force Grant AFOSR 72-2371 and in part by a Japanese Government Grant for Research at the Department of Electrical Engineering, University of Osaka, Suita, Japan.

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

The probabilistic search procedure presented here does not have these disadvantages. After proving the optimality of our procedure, both the convergence as it is defined for other random search procedures [14]–[16] and the convergences of state functions of interest in automata theory [1]–[4] are discussed. It is emphasized that the algorithm can easily cope with high noise and large strategy number situations. There is a great deal of freedom left to the designer within the boundaries dictated by the conditions of convergence. This freedom can be used to obtain fast-converging schemes.

It is indicated how the algorithm can be modified to operate in nonstationary environments. This modified procedure will be proved to be ε-optimal with respect to a certain function of the search parameters.

Later on, the organizational aspect of the search is briefly treated, and a specific design of the scheme is experimentally tested on the test problem of Shapiro and Narendra [6]. The rate of convergence for this scheme is considerably higher than for the SAVS, although further comparisons between the two techniques seem necessary.

II. The Probabilistic Search Procedure

The *environment* is characterized as follows. Consider the finite *set of strategies* $Z = \{z_1,\cdots,z_M\}$ and the set of probability measures μ_i with corresponding distribution function $F_i(\zeta)$, where $F_i(x) = P\{\zeta \le x \mid z_i\}$ is the probability of an environment's response ζ less than or equal to x, given that strategy z_i was applied to the environment. Define

$$Q(z_i) = E\{\zeta \mid z_i\} = \int x \, dF_i(x) \tag{1}$$

and assume, for simplicity, that

$$\begin{aligned} -\infty < Q_0^* = Q(z_1) < Q(z_2) \le Q(z_3) \le \cdots \le Q(z_M) \\ = Q_M^* < +\infty, \qquad Q(z_2) - Q(z_1) = D. \end{aligned} \tag{2}$$

It is desired to find the strategy with minimal $Q(z_i)$ while, at the same time, the average measured performance should converge in a certain fashion to $Q(z_1)$. The proposed procedure is iterative with iteration counter j. The *state* of the system (search procedure) is denoted by X_j and the *state space* by X. This state X_j completely determines the *set of selection probabilities* $\pi(X_j) = \{\pi_1(X_j),\cdots,\pi_M(X_j)\}$, where

$$\pi_i(X_j) = \alpha_j p_{0i} + (1-\alpha_j)p_{1i}(X_j), \qquad i = 1,\cdots,M \tag{3}$$

where $p_0 = \{p_{01},\cdots,p_{0M}\}$ is a set of fixed probabilities:

$$1 \ge p_{0i} > 0, \qquad \text{for all } i \qquad \sum_{i=1}^{M} p_{0i} = 1 \tag{4}$$

and $\{\alpha_j\}_{j\ge 0}$ is a sequence of numbers from $[0,1]$, $p_1(X_j)$ is a vector from $[0,1]^M$ with components $p_{1i}(X_j)$, $i = 1,\cdots,M$, that have a unit sum for all X_j:

$$\sum_{i=1}^{M} p_{1i}(X_j) = 1. \tag{5}$$

Thus, both p_0 and $p_1(X_j)$ are probability distributions on Z, and as a consequence, $\pi(X_j)$ also has all the properties of a probability distribution on Z. Notice here that the nature of $p_1(X_j)$ is arbitrary and does not play any role in establishing the convergence of the procedure. It will later be shown that $p_1(X_j)$ is important when it comes to accelerating the rate of convergence and making the scheme insensitive to search parameter selection.

The state X_j contains all the information concerning the history of the search up to the jth iteration that will be needed later on in the search process. We require, however, that X_j contain w_j (*basepoint* or best estimate of the optimal strategy up to the jth iteration) as a component, where, obviously, $w_j \in Z$.

The following procedure is a variant of the well-known random search algorithm [14]–[16].

1) X_0 is given to start the search process. At the jth iteration, we know X_j and thus w_j.
2) w_j is applied $\lambda_{Bj} \geq 1$ times to the environment, and λ_{Bj} i.i.d. (independent identically distributed) measurements are observed and averaged to yield an estimate ζ_j of $Q(w_j)$.
3) $w_{j+1}^* \in z$ is generated randomly according to the distribution $\pi(X_j)$ on Z.
4) w_{j+1}^* is applied $\lambda_{Tj} \geq 1$ times to the environment, and λ_{Tj} i.i.d. requirements ζ are observed and averaged to yield an estimate ζ_{j+1}^* of $Q(w_{j+1}^*)$.
5) X_j is updated through some rule T_{j+1}:

$$X_{j+1} = T_{j+1}(X_j, w_{j+1}^*, \zeta_j, \zeta_{j+1}^*, \cdots)$$

and we require only that w_j be updated as follows:

$$w_{j+1} = \begin{cases} w_{j+1}^*, & \text{if } \zeta_{j+1}^* < \zeta_j - \varepsilon_j \\ w_j, & \text{otherwise,} \end{cases} \tag{6}$$

where $\{\varepsilon_j\}_{j\geq 0}$ is a nonnegative number sequence.

First note that X_0 may be chosen arbitrarily since the convergence of the scheme does not rely on initial conditions. The next section will be devoted to the study of the asymptotical behavior of some functions of X_j (which are, of course, random variables). Of particular interest are

$$Q_j \triangleq Q(w_j) \tag{7}$$

$$V_j \triangleq \text{Ind}\,\{Q_j \leq Q(z_1)\}$$
$$= \text{Ind}\,\{w_j = z_1\} = \begin{cases} 1, & w_j = z_1 \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

$$R_j \triangleq \sum_{i=\zeta}^{M} \pi_i(X_j)Q(z_i) \tag{9}$$

$$S_j \triangleq \frac{\lambda_{Bj}Q_j + \lambda_{Tj}R_j}{\lambda_{Bj} + \lambda_{Tj}} \tag{10}$$

$$D_j \triangleq \frac{\lambda_{Bj}V_j + \lambda_{Tj}\pi_1(X_j)}{\lambda_{Bj} + \lambda_{Tj}}. \tag{11}$$

Q_j, the value of the performance index at the basepoint w_j, and V_j, the indicator function of the event $\{w_j = z_1\}$, are of interest in classical optimization where the way of obtaining estimates of the minimum is less important. State functions (9)–(11) correspond to random variables studied in automata theory. R_j is the expected value of $Q(w_{j+1}^*)$, S_j is the mean of the average measured performance, and D_j is the relative frequency of selection of the best strategy z_1 in $\lambda_{Bj} + \lambda_{Tj}$ "trials."

In analogy with the definition of expediency in automata theory [1], [2], [5], [10], we say that the search procedure is *expedient* if

$$\overline{\lim_{j\to\infty}}\, ES_j \leq \sum_{i=1}^{M} p_{0i}Q(z_i) \tag{12}$$

where $\overline{\lim}$ stands for lim sup. In the absence of any *a priori* information concerning the environment, it is reasonable to let $p_{0i} = (1/M)$, $i = 1,\cdots,M$.

The optimization scheme is said to be *optimal* if

$$\lim_{j\to\infty} ED_j = 1 \tag{13}$$

or, equivalently, since $D_j \in [0,1]$, if

$$D_j \to 1 \text{ in probability as } j \to \infty. \tag{14}$$

III. Theorem of Convergence

Environments are usually classified by the range of their responses. If $\zeta \in \{0,1\}$, then the environment is called a P-model environment. It is S-model if $\zeta \in [0,1]$. We would like to classify the environments as follows. An environment is of the L_{2r} type (where $1 \leq r \leq \infty$), if

$$\sup_{i=1,\cdots,M} \left(\int |\zeta - Q(z_i)|^{2r}\, dF_i(\zeta)\right)^{1/2r} = M_{2r} < \infty, \quad r < \infty$$
$$\sup_{i=1,\cdots,M} \text{ess sup}\, |\zeta - Q(z_i)| = M_\infty < \infty, \quad r = \infty \tag{15}$$

where the ess sup is with respect to $F_i(\zeta)$. Obviously, if an environment is of the L_{2r} type then it is of the L_{2s} type for all $1 \leq s \leq r$. P- and S-model environments are special cases of L_∞ type environments. If all the $\zeta - Q(z_i)$ are Gaussian, the environment is of the L_{2r} type for all $1 \leq r < \infty$ but is not of the L_∞ type. Because Gaussian environments play such an important role, we call them G-type environments and define

$$\sup_{i=1,\cdots,M} \int |\zeta - Q(z_i)|^2\, dF_i(\zeta) = M_G.$$

The main result is the following.

Theorem 1: Let (1) and (2) hold, $\{\alpha_j\}_{j\geq 0}$ be a number sequence from $[0,1]$, $\{\varepsilon_j\}_{j\geq 0}$ be a number sequence from $[0,\infty]$, $\{\lambda_{Bj}\}_{j\geq 0}$ and $\{\lambda_{Tj}\}_{j\geq 0}$ be integer sequences from $\{1,2,\cdots\}$, and let the state sequence $\mathscr{X} = \{X_j\}_{j\geq 0}$ be generated through procedure 1)–5) with $\pi(X_j)$ determined by (3)–(5). Further, let

$$\sum_{j=1}^{\infty} \alpha_j = \infty \tag{16}$$

and

$$\lim_{j\to\infty} \varepsilon_j = 0. \tag{17}$$

Then, i) if the environment is of the L_{2r} type $(1 \leq r < \infty)$,

$$\sum_{j=B_1}^{\infty} \lambda_{Bj}^{1/r} < \infty \qquad \sum_{j=1}^{\infty} \lambda_{Tj}^{1/r} < \infty \tag{18}$$

$\Rightarrow V_j \to 1$ with probability one as $j \to \infty$, and

$$\lim_{j\to\infty} \frac{1}{\alpha_j \lambda_{Bj}^r} = 0 \qquad \lim_{j\to\infty} \frac{1}{\alpha_j \lambda_{Tj}^r} = 0 \tag{19}$$

$\Rightarrow V_j \to 1$ in probability as $j \to \infty$. ii) If the environment is of the L_∞ type or G type,

$$\lim_{j\to\infty} \frac{\lambda_{Bj}}{\log j} = \infty \qquad \lim_{j\to\infty} \frac{\lambda_{Tj}}{\log j} = \infty \tag{20}$$

$\Rightarrow V_j \to 1$ with probability one as $j \to \infty$, and

$$\lim_{j\to\infty} \frac{\lambda_{Bj}}{\log \dfrac{1}{\alpha_j}} = \infty \qquad \lim_{j\to\infty} \frac{\lambda_{Tj}}{\log \dfrac{1}{\alpha_j}} = \infty \tag{21}$$

$\Rightarrow V_j \to 1$ in probability as $j \to \infty$.

Proof: The proof of Theorem 1 is based upon the theorems for the convergence of random processes proved by Braverman and Rozonoer [17]. The theorems we will use can be formulated as follows. If X_j is a sequence of random vectors on some probability space, if $U(X_j)$ is a requence of nonnegative random variables, and if $\{a_j\}_{j\geq 0}$ and $\{b_j\}_{j\geq 0}$ are number sequences from $[0,\infty)$ then if $EU(X_0)$ exists and

$$E\{U(X_{j+1})|X_j\} \leq U(X_j)(1 - a_j) + b_j, \qquad \text{for all } j \geq 0 \quad (22)$$

and if

$$\sum_{j=1}^{\infty} a_j = \infty \qquad \text{and} \qquad \lim_{j\to\infty} \frac{b_j}{a_j} = 0,$$

then $U(X_j) \to 0$ in probability as $j \to \infty$. If

$$\sum_{j=1}^{\infty} a_j = \infty \qquad \text{and} \qquad \sum_{j=1}^{\infty} b_j < \infty,$$

then $U(X_j) \to 0$ with probability one as $j \to \infty$. Suppose that $\varepsilon_j = 0$. We have,

$$\begin{aligned} E\{V_{j+1} \mid X_j\} \geq V_j \sum_{i=1}^{M} P\{w_{j+1} = w_j = z_1 \mid w_{j+1}^* = z_i,\\ w_j = z_1\}\pi_i(X_j)\\ + (1 - V_j) \min_{i=2,\cdots,M}\\ \cdot P\{w_{j+1} = w_{j+1}^* = z_1 \mid w_j = z_i,\\ w_{j+1}^* = z_1\}\pi_1(X_j) \quad (23)\end{aligned}$$

However, for all X_j, $\pi_i(X_j) \geq \alpha_j p_{0i}$ and

$$\begin{aligned} P\{w_{j+1} = w_j = z_1 \mid w_{j+1}^* = z_i, w_j = z_1\}\\ \geq P\left\{|\zeta_{j+1}^* - Q(z_i)| \leq \frac{D}{2} \middle| w_{j+1}^* = z_i\right\}\\ \cdot P\left\{|\zeta_j - Q(z_1)| \leq \frac{D}{2} \middle| w_j = z_1\right\} \quad (24)\end{aligned}$$

$$\begin{aligned} P\{w_{j+1} = w_{j+1}^* = z_1 \mid w_j = z_i, w_{j+1}^* = z_1\}\\ \geq P\left\{|\zeta_j - Q(z_i)| \leq \frac{D}{2} \middle| w_j = z_i\right\}\\ \cdot P\left\{|\zeta_{j+1}^* - Q(z_1)| \leq \frac{D}{2} \middle| w_{j+1}^* = z_1\right\} \quad (25)\end{aligned}$$

by virtue of (2), (6), and $\varepsilon_j = 0$.

Next, there exists a positive function $g(\cdot)$ such that, for all i,

$$P\left\{|\zeta_j - Q(z_i)| \geq \frac{D}{2} \middle| w_j = z_i\right\} \leq g(\lambda_{Bj}) \quad (26)$$

and

$$P\left\{|\zeta_{j+1}^* - Q(z_i)| \geq \frac{D}{2} \middle| w_{j+1}^* = z_i\right\} \leq g(\lambda_{Tj}) \quad (27)$$

where, if the environment is of the L_{2r} type ($1 \leq r \leq \infty$), by Markov's inequality [18] and Garsia's inequality for the expected value of the $2r$th moment of the sum of i.i.d. random variables [21]:

$$g(u) = \frac{2^r(2r)!}{r!} \cdot \frac{M_{2r}^{2r}}{\left(\frac{D}{2}\right)^{2r} \cdot u^r}, \qquad u \text{ integer}, \quad u \geq 1. \quad (28)$$

Also, for G-type environments, using Chernoff's bound [20] for Gaussian random variables:

$$g(u) = 2e^{-u(D/2)^2/2M_G}, \qquad u \text{ integer}, \quad u \geq 1. \quad (29)$$

If the environment is of the L_∞ type, we have by Hoeffding's inequality [19]:

$$g(u) = 2 \cdot e^{-2u \cdot (D/2)^2/(2M_\infty)^2}, \qquad u \text{ integer}, \quad u \geq 1. \quad (30)$$

Let $\gamma_j \triangleq g(\lambda_{Bj}) + g(\lambda_{Tj})$ and combine (22)–(27) to obtain

$$\begin{aligned} E\{(1 - V_{j+1}) \mid X_j\} &\leq (1 - V_j)(1 - \alpha_j p_{01}(1 + \gamma_j)) + \gamma_j\\ &\leq (1 - V_j)(1 - \alpha_j p_{01}) + \gamma_j \quad (31)\end{aligned}$$

so that, by $p_{01} > 0$, we need to ask that

$$\sum_{j=1}^{\infty} \alpha_j = \infty.$$

Furthermore, if $\gamma_j/\alpha_j \to 0$, then $V_j \to 1$ in probability as $j \to \infty$. If

$$\sum_{j=1}^{\infty} \gamma_j < \infty,$$

then $V_j \to 1$ with probability one as $j \to \infty$. With the proper substitution of γ_j (see (28)–(30)), conditions (18)–(21) are derived. If

$$\lim_{j\to\infty} \varepsilon_j = 0$$

then $\varepsilon_j < D/2$, for all j large enough. Clearly, (31) still holds for all such j if in the definition of γ_j, D is replaced by $D/2$. This completes the proof.

Remark 1: If $\alpha_j \geq \alpha > 0$ for all j, conditions (19) and (21) are very weak. In particular, for all L_2-type environments, (19) implies that λ_{Bj} and λ_{Tj} should diverge at any rate, however low. The condition that an environment is of the L_2 type is, in practice, always fulfilled because most types of noise on responses from real systems have bounded variance.

For L_∞- or G-type environments, if the rate of increase of λ_{Bj} and λ_{Tj} is faster than logarithmic, V_j converges to 1 with probability one (by (20)), even if $\alpha_j \to 0$ as $j \to \infty$.

Remark 2: It is easy to see that convergence in probability to Q_0^* and convergence of the mean to Q_0^* are equivalent for Q_j, R_j, and S_j, all of which take values in $[Q_0^*, Q_M^*]$. Convergence in probability and of the mean to 1 are equivalent for the random variables V_j and D_j, both of which take values in $[0,1]$. Also, $V_j \to 1$ in probability (with probability one) $\Rightarrow Q_j \to Q_0^*$ in probability (with probability one) since $Q_j \leq Q_0^* \cdot V_j + Q_M^* \cdot (1 - V_j)$.

Finally, $V_j \to 1$ in probability (with probability one) and

$$\lim_{j\to\infty} \frac{\lambda_{Tj}}{\lambda_{Bj} + \lambda_{Tj}} = 0 \quad (32)$$

together imply that $D_j \to 1$ and $S_j \to Q_0^*$ in probability (with probability one) since

$$D_j \geq \frac{\lambda_{Bj}}{\lambda_{Bj} + \lambda_{Tj}} \cdot V_j$$

and

$$S_j \leq Q_0^* \cdot V_j \cdot \frac{\lambda_{Bj}}{\lambda_{Bj} + \lambda_{Tj}} + Q_M^* \cdot \left(1 - \frac{\lambda_{Bj}}{\lambda_{Bj} + \lambda_{Tj}} \cdot V_j\right).$$

Remark 3: The reader may wonder why one does not let $\varepsilon_j \equiv 0$ for all j. Experience with random search algorithms has led several authors [14], [15] to believe that a nonzero ε_j keeps the algorithm from changing the basepoint too frequently and too carelessly. Only when λ_{Bj} and λ_{Tj} are large enough so that ζ_j and ζ_{j+1}^* are good estimates of $Q(w_j)$ and $Q(w_{j+1}^*)$ can we let ε_j be small without having to fear a wrong decision in (6).

IV. Epsilon-Optimality

In nonstationary environments, the same procedure with constant parameters $\alpha_j = \alpha$, $\varepsilon_j = \varepsilon$, $\lambda_{Bj} = \lambda_B$ and $\lambda_{Tj} = \lambda_T$ is of definite interest. Without pretending that this constant parameter procedure will be powerful in nonstationary environments, we will just show that the so-obtained algorithm is ε-optimal in stationary environments.

We will say that the search procedure is *ε-optimal* if, for all $\eta > 0$, we can choose the search parameters (here: α, ε, λ_B, λ_T) in such a way that

$$\underline{\lim}\, ED_j \geq 1 - \eta, \qquad \varepsilon\text{-optimality for } D_j \tag{33}$$

or

$$\underline{\lim}\, EV_j \geq 1 - \eta, \qquad \varepsilon\text{-optimality for } V_j \tag{34}$$

where $\underline{\lim}$ stands for $\lim\inf_{j\to\infty}$. It has been proven by Sawaragi and Baba [12] that the L_{R-I} (linear reward-inaction) SAVS is ε-optimal (for a definition, see [8], [12]) for $\pi_1(X_j)$. However, in order to account for the λ_B and λ_T measurements made at each iteration, we needed this broader definition.

Theorem 2: If (1) and (2) hold, $\alpha \in [0,1]$, $\varepsilon \geq 0$, $\lambda_T \geq 1$, $\lambda_B \geq 1$, and if the state sequence $\mathscr{X}$ is generated through procedure 1)–5) with $\pi(X_j)$ determined by (3)–(5) and if the environment is at least of the L_2 type, then the presented search procedure is ε-optimal both for D_j and V_j.

Proof: Using (31) and (28) with $r = 1$ and $M_2{}^2 = M_G$ and letting $\varepsilon = 0$:

$$E\{(1 - V_{j+1}) \mid X_j\} \leq (1 - V_j)(1 - \alpha p_{01}) + \frac{4M_G}{(D/2)^2} \cdot \left(\frac{1}{\lambda_B} + \frac{1}{\lambda_T}\right).$$

Let

$$C \triangleq \frac{4M_G}{(D/2)^2}.$$

Taking expectations at both sides gives

$$E\{1 - V_{j+1}\} \leq (1 - \alpha \cdot p_{01}) \cdot E\{1 - V_j\} + C\left(\frac{1}{\lambda_B} + \frac{1}{\lambda_T}\right).$$

Recursive computation yields

$$E\{1 - V_{j+1}\} \leq E\{1 - V_0\}(1 - \alpha p_{01})^j \cdot \left(1 - \alpha p_{01} + C\left(\frac{1}{\lambda_B} + \frac{1}{\lambda_T}\right)\right) + \frac{C}{\alpha p_{01}}\left(\frac{1}{\lambda_B} + \frac{1}{\lambda_T}\right)(1 - (1 - \alpha \cdot p_{01})^{j+1}).$$

Because $E\{1 - V_0\} \geq 0$ and $(1 - \alpha p_{01})^j \to 0$ as $j \to \infty$,

$$\overline{\lim_{j\to\infty}}\, E\{1 - V_{j+1}\} \leq \frac{C}{\alpha p_{01}}\left(\frac{1}{\lambda_B} + \frac{1}{\lambda_T}\right).$$

The right side can be made smaller than η by taking λ_B and λ_T large enough. From definition (11) and $\pi_1(X_j) \geq 0$:

$$E\{1 - D_j\} \leq \frac{\lambda_T}{\lambda_B + \lambda_T} + \frac{\lambda_B}{\lambda_B + \lambda_T} E\{1 - V_j\}$$

so that

$$\overline{\lim_{j\to\infty}}\, E\{1 - D_j\} \leq \frac{\lambda_T}{\lambda_B + \lambda_T} + \frac{C}{\alpha \cdot p_{01}} \cdot \left(\frac{1}{\lambda_B} + \frac{1}{\lambda_T}\right)$$

which can be made smaller than η by choosing λ_B and λ_T large enough and $\lambda_T/(\lambda_B + \lambda_T)$ small enough.

V. Organization of The Search

It is still an open problem whether optimal number sequences can be found for the search parameters or not (as in stochastic approximation algorithms). Besides this, there is the organizational aspect of the search parameter selection, which usually involves heuristics based upon the experience of the designer. Some *a priori* information can always help in selecting suitable sequences $\{\lambda_{Bj}\}$, $\{\lambda_{Tj}\}$, $\{\alpha_j\}$ and $\{\varepsilon_j\}$. Notice also that it is possible to adapt the search parameters within boundaries that are number sequences satisfying the conditions of Theorem 1.

Furthermore, there is the still undiscussed distribution $p_1(X_j)$ (3)–(5) which is very important because $\pi(X_j)$ is very nearly equal to $p_1(X_j)$ for large j if $\alpha_j \to 0$ as $j \to \infty$. The following distribution is proposed: a record is kept of all the observations made in the past with each strategy. For z_i, let $v_{i,j}$ denote the number of measurements observed up to iteration j after z_i was applied to the environment. Let $\mu_{i,j}$ be the mean over these $v_{i,j}$ measurements and let $\tau_{i,j}$ be the quadratic mean of these measurements. μ_j, τ_j and v_j are M-dimensional vectors grouping $\mu_{i,j}$, $\tau_{i,j}$, and $v_{i,j}$ for $i = 1,\cdots,M$. The state X_j now contains quite a lot of information about the random environment and the history of the search because $X_j = \{w_j,\mu_j,\tau_j,v_j,\cdots\}$. Luckily, we know that μ_j, τ_j, and v_j can be recursively updated as new measurements come in so that the observed environment responses need not be stored.

Let

$$s^2(z_i) = E\{\zeta^2 \mid z_i\} = \int \zeta^2 \, dF_i(\zeta)$$

so that

$$\int |\zeta - Q(z_i)|^2 \, dF_i(\zeta) = s^2(z_i) - Q^2(z_i).$$

Since

$$\sum_{j=1}^{\infty} \alpha_j = \infty,$$

$v_{i,j} \to \infty$ with probability one, for all $i = 1,\cdots,M$. Consequently, $\mu_{i,j}$ will approximate $Q(z_i)$ and $\tau_{i,j}$ will approximate $s^2(z_i)$ if the environment is at least of the L_2 type (so that $s^2(z_i) < \infty$, for all $i = 1,\cdots,M$).

Suppose that $Q(z_i)$ and $s^2(z_i)$ were known and that one would take $v_{i,j}$ i.i.d. measurements with z_i and denote the average by ζ_j. Using Chebyshev's inequality:

$$P\{\zeta_j \leq Q_0{}^*\} \leq \frac{s^2(z_i) - Q^2(z_i)}{v_{i,j} \cdot (Q(z_i) - Q_0{}^*)^2}. \tag{35}$$

The right side of (35) is not known but can be estimated using the data available in X_j. The estimate is denoted by $\beta_{i,j}$:

$$\beta_{i,j} \triangleq \begin{cases} \dfrac{\tau_{i,j} - \mu_{i,j}{}^2}{v_{i,j} \cdot (\mu_{i,j} - \min_{i'} \mu_{i',j})^2}, & \text{if } \mu_{i,j} > \min_{i'} \mu_{i',j} \\ \infty, & \text{otherwise.} \end{cases} \tag{36}$$

$\beta_{i,j}$ is larger if a) $v_{i,j}$ is smaller (z_i is not very frequently used up to the jth iteration), b) $\mu_{i,j}$ is smaller relative to the $\mu_{i',j}$, $i' \neq i$. (z_i is a promising strategy with high probability that the corresponding $Q(z_i)$ is small), c) $\tau_{i,j} - \mu_{i,j}{}^2$ is larger (which would indicate that $s^2(z_i)$ is large and that more sampling with z_i is needed to obtain a low variance on the estimate $\mu_{i',j}$ of $Q(z_i)$).

This shows that $\beta_{i,j}$ is, in fact, proportional to the need of selecting strategy z_i. Define $p_1(X_j)$ by

$$p_{1i}(X_j) = \left(\frac{\gamma}{\gamma + 1/\beta_{i,j}}\right) \Big/ \left[\sum_{i'=1}^{M} \left(\frac{\gamma}{\gamma + 1/\beta_{i',j}}\right)\right], \quad i = 1,\cdots,M \tag{37}$$

where $\gamma > 0$ is to be chosen by the designer. It was mentioned that $\mu_{i,j} \to Q(z_i)$ and $\tau_{i,j} \to s^2(z_i)$ at least in probability so that $\beta_{1,j} \to \infty$ and $\beta_{i,j} \to 0$ for $i \geq 2$ in probability. Therefore,

$$p_{1i}(X_j) \to \begin{cases} 1, & i = 1 \\ 0, & i \geq 2 \end{cases}$$

in probability as $j \to \infty$. The exact result is the following theorem.

Theorem 3: Let (1) and (2) hold, $\{\alpha_j\}_{j\geq 0}$ be a number sequence from [0,1] and $\{\lambda_{Tj}\}_{j\geq 0}$ be an integer sequence from $[1,\infty)$, and let $\pi(X_j)$ be determined by (3)–(5) and $p_1(X_j)$ be determined by (35)–(37). Let the environment be at least of the L_2 type, let

$$\sum_{j=1}^{\infty} \alpha_j \cdot \lambda_{Tj} = \infty, \tag{38}$$

and let there exist a $B < \infty$ such that

$$\sup_j \frac{\sum_{k=1}^{j} \alpha_k \cdot \lambda_{Tk}^2}{\left(\sum_{k=1}^{j} \alpha_k \cdot \lambda_{Tk}\right)^2} \leq B < \infty. \tag{39}$$

Let the state sequence $\mathscr{X}$ be determined *either* by procedure 1)–5) (in which case we need to ask that $\{\lambda_{Bj}\}_{j\geq 0}$ is an integer sequence from $[1,\infty)$ and $\{\varepsilon_j\}_{j\geq 0}$ is a nonnegative number sequence) *or* by step 4) alone (in which case one needs not store w_j, and the algorithm consists of updating $p_1(X_j)$ through (35)–(37)). Then

$$p_{1i}(X_j) \to \begin{cases} 1, & \text{in probability as } j \to \infty,\ i = 1 \\ 0, & \text{in probability as } j \to \infty,\ i \geq 2. \end{cases}$$

The proof is given in the Appendix.

Corollary: If, in addition to the requirements of the theorem, $\lim_{j\to\infty} \alpha_j = 0$, then $\pi_1(X_j) \to 1$ in probability as $j \to \infty$ and $R_j \to Q_0^*$ in probability as $j \to \infty$.

Remark 1: There are no restrictions on λ_{Bj} and ε_j if procedure 4) is followed. One can thus as well let λ_{Tj} be constant, for instance 1, for all j, in which case the procedure looks very much like an automaton where one observation is made per iteration.

Remark 2: Condition (39) is, for instance, fulfilled if $B < \infty$ is such that

$$\sup_j \frac{\sup_{k\leq j} \lambda_{Tk}}{\sum_{k=1}^{j} \alpha_k \cdot \lambda_{Tk}} \leq B$$

and does not allow the sequence $\{\lambda_{Tj}\}$ to be too oscillatory with high "peaks," thus increasing the variances of the $v_{i,j}$ too rapidly relative to the increase in $Ev_{i,j}$.

VI. Experiments

The presented algorithm with given choice of $p_1(X_j)$ (35)–(37) is used in the test problem of Shapiro and Narendra [6] where $M = 10$ and $\{Q(z_1),\cdots,Q(z_{10})\} = \{-5.6,-5.5,-5.3,-5.3,-5.1,-5.1,-5.1,-5.0,-4.9,-4.9\}$. Thus $D = 0.1$, and if $p_0 = \{0.1,\cdots,0.1\}$,

$$\sum_{i=1}^{M} p_{0i} \cdot Q(z_i) = -5.18.$$

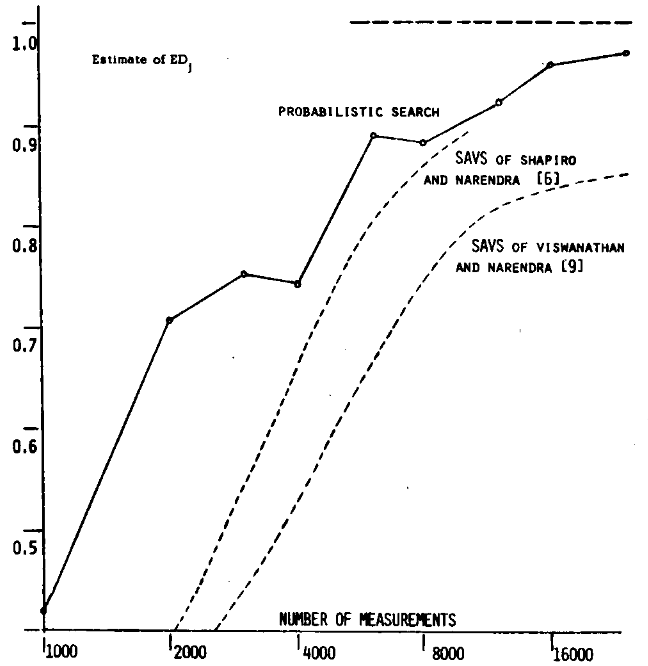

Fig. 1. Estimate of ED_j versus number of measurements for test problem of Shapiro and Narendra [6].

The environment is of the L_∞ type since $F_i(\zeta)$ is the uniform distribution function in $[Q(z_i) - 2, Q(z_i) + 2]$.

One measure of the difficulty of a problem is the ratio $M \cdot (M_G/D^2)$ and environments with ratios below five can be considered as relatively "easy" environments. In Narendra's test problem, however, $M \cdot (M_G/D^2)$ roughly equals 1300. The following number sequences were used:

$$\lambda_{Bj} = \max\{5;(\lambda_0 \cdot j)^{1.3}\}$$

$$\lambda_{Tj} = \max\{5;(\lambda_0 \cdot j)^{0.9}\}$$

with $\lambda_0 = 4$, $\varepsilon_j = 0.08 < D$ (which is, in fact, sufficient to make Theorem 1 work. If D is unknown, however, it is necessary to require that $\lim_{j\to\infty} \varepsilon_j = 0$), $\alpha_j = (0.2/j)^a (a \in [0,1])$, and $p_1(X_j)$ is defined by (35)–(37) with $\gamma = 1$. To start the search, $n_0 = 100$ measurements are made with each strategy $z_i \in Z$ (thus let $v_{i,0} = n_0$, $i = 1,\cdots,M$ and $w_0 = z_{i*}$ where i^* is defined by $\mu_{i*,0} = \min_i \mu_{i,0}$).

The curves of Fig. 1 give 50-run averages of D_j as a function of L_j, the number of measurements up to iteration j, i.e.,

$$\sum_{i=0}^{j} (\lambda_{Bi} + \lambda_{Ti}) + M \cdot n_0,$$

25 runs of which were with $a = 0.8$ and 25 runs with $a = 1.0$. The dotted lines are the results obtained by Shapiro and Narendra [6] and Viswanathan and Narendra [9] for the same test problem with SAVS schemes that are adapted for use in general environments. For the SAVS, where $\lambda_{Bj} = 0$, $\lambda_{Tj} = 1$, and the algorithm reduces to updating $\pi(X_j)$ after each observation, D_j clearly equals $\pi_1(X_j)$.

Considering that the abscis scale is logarithmic, a comfortable improvement in the rate of convergence is obtained as is seen from Fig. 1.

To demonstrate the relative insensitivity with respect to the selection of gain factors such as λ_0, the same experiment is repeated, and D_j is averaged over 50 runs, 25 with $a = 0.8$ and 25 with $a = 1.0$. These averages are depicted as a function of

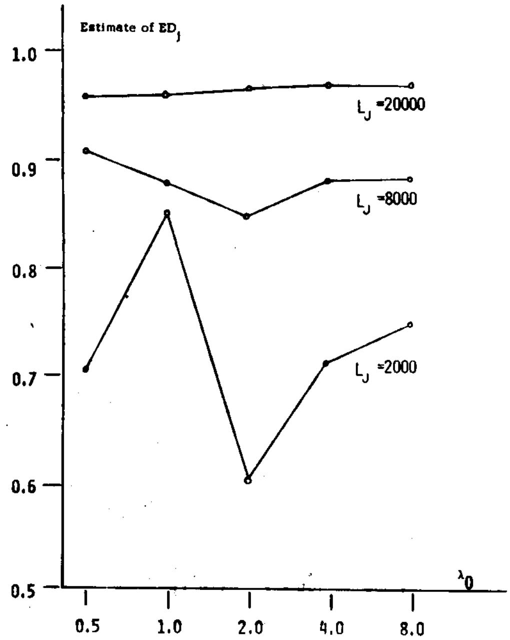

Fig. 2. Estimate of ED_J versus search parameter λ_0 for test problem of Shapiro and Narendra [6].

λ_0 for $L_J = 2000$, $L_J = 8000$, and $L_J = 20\,000$. λ_0 varies from 0.5 to 8.0. The curves of Fig. 2 level off as the number of observations (L_j) increases. This seems to indicate that it is not so important if a few iterations are made with high sampling rates λ_{Bj} and λ_{Tj} or if many iterations are made with low sampling rates. We indicated already that $\pi_1(X_j) \to 1$ in probability and that $R_j \to Q_0^*$ in probability as $j \to \infty$. For large j, w_{j+1}^* has high probability to be equal to z_1, and since most of the time either $w_j = z_1$ or $w_{j+1}^* = z_1$ or both, it becomes less relevant whether the decision (6) is based upon large or small λ_{Bj}, λ_{Tj}. It is thus the special choice of $p_1(X_j)$ which is the predominant factor for insuring a high rate of convergence and low sensitivity with respect to the search parameter selection. Notice finally that as for most probabilistic global search procedures [15]–[16], there is no sensitivity regarding initial conditions.

Although all these properties make the proposed method very attractive, further research is still desired to make the scheme completely self-organizing. This involves the development of higher-level learning or adaptation of λ_{Bj}, λ_{Tj}, α_j, and ε_j without losing the nice convergence properties obtained in this paper.

VII. Conclusion

It is shown that the problem of the selection of the best strategy in L_2-type random environments can also be solved through probabilistic search procedures. The asymptotic optimality of the method, proved in Theorem 1, is often of more theoretical than practical value, however. This practical barrier has been overcome by the proposed almost completely self-organizing probabilistic search scheme featuring insensitivity regarding initial conditions and search parameter selection.

The first experimental comparisons between probabilistic search and SAVS seem to indicate that, at least in stationary and high-noise environments, probabilistic search has a superior rate of convergence. In nonstationary environments, the proposed fixed-parameter version of the algorithm, proved to be ε-optimal in Theorem 2, should be compared with some of the SAVS that are powerful in such environments.

Appendix

Proof of Theorem 3

Clearly, it suffices to show that $p_{1i}(X_j) \to 0$ in probability as $j \to \infty$ for all $i \geq 2$. By (37):

$$p_{1i}(X_j) \leq \frac{\gamma + 1/\beta_{1,j}}{\gamma + 1/\beta_{i,j}} \leq \beta_{i,j}\left(\gamma + \frac{1}{\beta_{1,j}}\right). \tag{40}$$

We show that $\beta_{1,j} \to \infty$ in probability and $\beta_{i,j} \to 0$ in probability as $j \to \infty$.

First, let $P_{\min} \triangleq \min\{p_{01}, \cdots, p_{0M}\} > 0$, and let $\eta > 0$ be given. We shall first show that for all j large enough,

$$P\{\beta_{1,j} = \infty\} \geq 1 - \eta.$$

First, choose $L \in (0,\infty)$ so large that $(L-1)^2 \geq [2M(2+B)]/\eta$ and define

$$N_j = L \cdot \sum_{k=1}^{j} \alpha_k \cdot \lambda_{Tk}.$$

Notice that (38) implies that $\lim_{j\to\infty} N_j = \infty$. Define further the event

$$A_j = \{\min_{i=1,\cdots,M} \nu_{i,j} \geq N_j\}. \tag{41}$$

Thus $P\{\beta_{1,j} = \infty\} \geq P\{A_j\}$.

$$P\{\mu_{1,j} = \min_i \mu_{i,j} \mid A_j\}$$

where, if $\hat{\mu}_{i,n}$ is the average of n i.i.d. random variables with cumulative distribution function (cdf) $F_i(\zeta)$ and mean $Q(z_i)$,

$$\begin{aligned} P\{\mu_{1,j} > \min_i \mu_{i,j} \mid A_j\} &\leq \sum_{i=1}^{M} P\left\{|\mu_{i,j} - Q(z_i)| > \frac{D}{2} \,\middle|\, A_j\right\} \\ &\leq \sum_{i=1}^{M} P\left\{\sup_{n\geq N_j} |\hat{\mu}_{i,n} - Q(z_i)| > \frac{D}{2}\right\} \\ &< \frac{\eta}{2} \end{aligned} \tag{42}$$

for all j large enough by the strong law of large numbers [18] and the fact that $N_j \to \infty$ as $j \to \infty$. Let $\nu_{i,j}^*$ denote the fraction of the $\nu_{i,j}$ observations made with z_i when z_i was generated in step 3) and with probability $p_{0i} \cdot \alpha_k$ in (3).

$$P\{A_j^c\} \leq \sum_{i=1}^{M} P\{\nu_{i,j}^* < N_j\} \leq \sum_{i=1}^{M} \frac{E\nu_{i,j}^{*\,2}}{(N_j - E\nu_{i,j}^*)^2}.$$

Noting that

$$E\nu_{i,j}^{*\,2} \leq \sum_{k=1}^{j} \alpha_k \lambda_{Tk}^2 + 2\left(\sum_{k=1}^{j} \alpha_k \cdot \lambda_{Tk}\right)^2$$

and

$$N_j - E\nu_{i,j}^* \geq (L-1)\sum_{k=1}^{j} \alpha_k \cdot \lambda_{Tk},$$

$$\begin{aligned} P\{A_j^c\} &\leq \frac{M}{(L-1)^2}\cdot\left(2 + \frac{\sum_{k=1}^{j}\alpha_k \cdot \lambda_{Tk}^2}{\left(\sum_{k=1}^{j}\alpha_k\cdot\lambda_{Tk}\right)^2}\right) \\ &\leq \frac{M(2+B)}{(L-1)^2} \\ &< \frac{\eta}{2} \end{aligned}$$

by choice of L. Therefore,

$$P\{\beta_{1,j} = \infty\} \geq (1 - P\{A_j^c\}) \cdot (1 - P\{\mu_{1,j} > \min_i \mu_{i,j} \mid A_j\}) \geq \left(1 - \frac{\eta}{2}\right)^2 > 1 - \eta \tag{43}$$

for all j large enough.

Next, we show that, given $\eta > 0$ arbitrary, and $\delta > 0$ arbitrary, $P\{\beta_{i,j} \leq \delta\} > 1 - \eta$, for all j large enough. Without loss of generality, let $i = 2$. Let N_j be defined as before and let A_j be given by (41) and

$$B_j \triangleq \left\{\bigcap_{i=1}^{M} |\mu_{i,j} - Q(z_i)| < \frac{D}{2}\right\}.$$

Then

$$P\{\beta_{i,j} \leq \delta\} \geq P\{A_j\}P\{B_j \mid A_j\} \cdot P\left\{|\tau_{2,j} - \mu_{2,j}^2| < \left(\frac{D}{2}\right)^2 \cdot N_j \cdot \delta \mid A_j B_j\right\}$$

$P\{A_j\} \geq 1 - (\eta/2)$ as shown before and using the same strong law of large numbers argument, $P\{B_j \mid A_j\} \geq 1 - (\eta/4)$, for all j large enough. Next,

$$\begin{aligned} P&\left\{\tau_{2,j} - \mu_{2,j}^2 \geq \left(\frac{D}{2}\right)^2 \cdot N_j\delta \mid A_j B_j\right\} \\ &\leq P\left\{|\tau_{2,j} - s^2(z_2)| \geq \left(\frac{D}{2}\right)^2 N_j \cdot \frac{\delta}{2} \,\middle|\, A_j B_j\right\} \\ &\quad + P\left\{s^2(z_2) \geq \left(\frac{D}{2}\right)^2 \cdot N_j \cdot \frac{\delta}{2} \,\middle|\, A_j B_j\right\} \end{aligned} \tag{44}$$

where $s^2(z_2) \triangleq \int \zeta^2 \cdot dF_i(\zeta)$. The last term of (44) is 0 for large j since $N_j \to \infty$ and $s^2(z_2) < \infty$ for L_2-type environments. Let

$$C_j \triangleq \left\{|\tau_{2,j} - s^2(z_2)| \geq \left(\frac{D}{2}\right)^2 \cdot N_j \cdot \frac{\delta}{2}\right\}$$

and note that for all j large enough:

$$P\{C_j \mid A_j \cdot B_j\} = \frac{P\{C_j B_j \mid A_j\}}{P\{B_j\}} \leq \frac{P\{C_j \mid A_j\}}{P\{B_j A_j\}} \leq \frac{P\{C_j \mid A_j\}}{\left(1 - \frac{\eta}{2}\right)\left(1 - \frac{\eta}{4}\right)},$$

and if $\hat{\tau}_{2,n}$ is the average of n i.i.d. random variables distributed as Y^2 where Y has distribution $F_i(\zeta)$ and Y^2 has mean $s^2(z_2)$,

$$P\{C_j \mid A_j\} \leq \left\{\sup_{n \geq N_j} |\hat{\tau}_{2,n} - s^2(z_2)| \geq \left(\frac{D}{2}\right)^2 \cdot N_j \cdot \frac{\delta}{2}\right\}.$$

Obviously, since $s^2(z_2) < \infty$, the strong law of large numbers applies to the sequence $\{\hat{\tau}_{2,n}\}_{n \geq 1}$, and since $N_j \to \infty$ as $j \to \infty$, $P\{C_j \mid A_j\} < (\eta/4)(1 - (\eta/2))(1 - (\eta/4))$ for all j large enough. Thus, for all j large enough,

$$P\{\beta_{2,j} \leq \delta\} \geq \left(1 - \frac{\eta}{2}\right)\left(1 - \frac{\eta}{4}\right)\left(1 - \frac{\eta}{4}\right) > 1 - \eta.$$

This completes the proof of Theorem 3.

References

[1] M. L. Tsetlin, "On the behaviour of finite automata in random media," *Aut. and Remote Contr.*, vol. 22, no. 10, pp. 1345–1354, 1961.

[2] V. I. Varshavskii and I. P. Vorontsova, "On the behaviour of stochastic automata with a variable structure," *Aut. and Remote Contr.*, vol. 24, no. 3, pp. 353–360, 1963.

[3] K. S. Fu and R. W. Maclaren, "An application of stochastic automata to the synthesis of learning systems," Purdue Univ., West Lafayette, IN, Tech. Rep. TR-EE-6517, 1965.

[4] K. S. Fu and G. J. McMurtry, "A study of stochastic automata as models of adaptive and learning controllers," Purdue Univ., West Lafayette, IN, Tech. Rep. TR-EE-65-08, 1965.

[5] B. Chandrasekaran and D. W. C. Shen, "On expediency and convergence in variable structure automata," *IEEE Trans. Syst. Sci. Cybern.*, vol. SSC-4, pp. 52–60, Mar. 1968.

[6] I. J. Shapiro and K. S. Narendra, "Use of stochastic automata for parameter self-optimization with multimodal performance criteria," *IEEE Trans. Syst. Sci. Cybern.*, vol. SSC-5, pp. 352–360, July 1969.

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

[8] R. Viswanathan and K. S. Narendra, "A note on the linear reinforcement scheme for variable structure stochastic automata," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-2, pp. 292–294, Apr. 1972.

[9] ——, "Stochastic automata models with applications to learning systems," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-3, pp. 107–111, Jan. 1973.

[10] S. Lakshmivarahan and M. A. L. Thathachar, "Absolutely expedient learning algorithms for stochastic automata," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-3, pp. 281–286, May 1973.

[11] L. G. Mason, "An optimal learning algorithm for *S*-model environments," *IEEE Trans. Aut. Contr.*, vol. AC-18, pp. 493–496, Sept. 1973.

[12] Y. Sawaragi and N. Baba, "A note on the learning behaviour of variable structure stochastic automata," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-3, pp. 644–649, Nov. 1973.

[13] M. F. Norman, *Markov Processes and Learning Models.* New York: Academic, 1963.

[14] L. S. Gurin, "Random search in the presence of noise," *Engineering Cybern.*, vol. 4, no. 3, pp. 252–260, 1966.

[15] L. D. Cockrell and K. S. Fu, "On search techniques in adaptive systems," Purdue Univ., West Lafayette, IN, Tech. Rep. TR-EE-70-01, 1970.

[16] G. J. McMurtry, "Adaptive optimization procedures," in *Adaptive, Learning and Pattern Recognition Systems*, J. M. Mendel and K. S. Fu, Eds. New York: Academic, 1970.

[17] E. M. Braverman and L. I. Rozonoer, "Convergence of random processes in learning machines theory, Part 1," *Aut. and Remote Contr.*, vol. 30, no. 1, pp. 57–77, 1969.

[18] M. Loeve, *Probability Theory.* Princeton, NJ: Van Nostrand, 1963.

[19] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *J. Amer. Statist. Assoc.*, vol. 58, no. 1, pp. 13–30, 1963.

[20] A. M. Garsia, *Topics in Almost Everywhere Convergence.* Chicago: Markham, 1970.

[21] H. Chernoff, "A measure of asymptotic efficiency for test of a hypothesis based on the sum of observations," *Ann. Math. Stat.*, vol. 23, pp. 493–507, 1952.