Learning Automata—A Survey

KUMPATI S. NARENDRA, SENIOR MEMBER, IEEE, AND M. A. L. THATHACHAR

***Abstract*—Stochastic automata operating in an unknown random environment have been proposed earlier as models of learning. These automata update their action probabilities in accordance with the inputs received from the environment and can improve their own performance during operation. In this context they are referred to as learning automata. A survey of the available results in the area of learning automata has been attempted in this paper. Attention has been focused on the norms of behavior of learning automata, issues in the design of updating schemes, convergence of the action probabilities, and interaction of several automata. Utilization of learning automata in parameter optimization and hypothesis testing is discussed, and potential areas of application are suggested.**

I. Introduction

In CLASSICAL deterministic control theory, the control of a process is always preceded by complete knowledge of the characteristics of the process; the mathematical description of the process is assumed to be known, and the inputs to the process are deterministic functions of time. Later developments in stochastic control theory took into account uncertainties that might be present in the process; stochastic control was effected by assuming that the probabilistic characteristics of the uncertainties are known. Frequently, the uncertainties are of a higher order, and even the probabilistic characteristics such as the distribution functions may not be completely known. It is then necessary to make observations on the process as it is in operation and gain further knowledge of the process. In other words, a distinctive feature of such problems is that there is little *a priori* information, and additional information is to be acquired on line. One viewpoint is to regard these as problems in learning.

Learning is defined as any relatively permanent change in behavior resulting from past experience, and a learning system is characterized by its ability to improve its behavior with time, in some sense tending towards an ultimate goal. In mathematical psychology, models of learning systems [GB1], [GL1] have been developed to explain behavior patterns among living organisms. These models in turn have lately been adapted to synthesize engineering systems, which can be considered to show learning behavior. Tsypkin [GT1] has recently argued that seemingly diverse problems in pattern recognition, control, identification, filtering, etc. can be treated in a unified manner as problems in learning using probabilistic iterative methods.

Viewed in a purely mathematical context the goal of a learning system is the optimization of a functional not known explicitly, as, for example, the mathematical expectation of a random functional with a probability distribution function not known in advance. An approach that has been used in the past is to reduce the problem to the determination of an optimal set of parameters and then apply stochastic hillclimbing techniques [GT1]. An alternative approach gaining attention recently is to regard the problem as one of finding an optimal action out of a set of allowable actions and to achieve this using stochastic automata [LN2]. The following example of the learning process of a student with a probabilistic teacher illustrates the automaton approach.

Consider a student–teacher pair. A question is posed to the student, and a finite set of alternative answers is provided. The student can select one of the alternatives, following which the teacher responds in a binary manner indicating whether the selected answer is right or wrong. The teacher is, however, probabilistic—there is a nonzero probability of eliciting either of the two responses for any of the answers selected by the student. The saving feature of the situation is that it is known that the teacher's negative responses have the least probability for the correct answer. Under these circumstances the interest is in finding the manner in which the student should plan a choice of a sequence of alternatives and process the information obtained from the teacher so that he learns the correct answer.

In stochastic automata models the stochastic automaton corresponds to the student, and the random environment in which it operates represents the probabilistic teacher. The actions (or states) of the stochastic automaton are the various alternative answers that are provided. The responses of the environment for a particular action of the stochastic automaton are the teacher's probabilistic responses. The problem is to obtain the optimal action that corresponds to the correct answer.

The stochastic automaton attempts a solution of this problem as follows. To start with, no information as to which one is the optimal action is assumed, and equal

Manuscript received January 15, 1974; revised February 13, 1974. This work was supported by the National Science Foundation under Grant GK-20580.

K. S. Narendra is with the Becton Center, Yale University, New Haven, Conn.

M. A. L. Thathachar is with the Becton Center, Yale University, New Haven, Conn., on leave from the Indian Institute of Science, Bangalore, India.

probabilities are attached to all the actions. One action is selected at random, the response of the environment to this action is observed, and based on this response the action probabilities are changed. Now a new action is selected according to the updated action probabilities, and the procedure is repeated. A stochastic automaton acting in this manner to improve its performance is referred to as a *learning automaton* in this paper.

Stochastic hillclimbing methods (such as stochastic approximation) and stochastic automata methods represent two distinct approaches to the learning problem. Though both approaches involve iterative procedures, updating at every stage is done in the parameter space in the first method and probability space in the second. It is, of course, possible that they lead to equivalent descriptions in some examples. The automata methods have two distinct advantages over stochastic hillclimbing methods in that the action space need not be a metric space (i.e., no concept of neighborhood is needed), and since at every stage any element of the action set can be chosen, global rather than local optimum can be obtained.

Experimental simulation of automata methods carried out during the last few years has indicated the feasibility of the automaton approach in the solution of interesting examples in parameter optimization, hypothesis testing, and game theory. The automaton approach also appears appropriate in the study of hierarchical systems and in tackling certain nonstationary optimization problems. Furthermore, several other avenues to learning can be interpreted as iterative procedures in the probability space, and the learning automaton provides a natural mathematical model for such situations and serves as a unifying theme among diverse techniques [GM3].

Previous studies on learning automata have led to a certain understanding of the basic issues involved and have provided guidelines for the design of algorithms. An appreciation of the fundamental problems in the field has also taken place. It appears that research in this area has reached a stage where the power and applicability of the approach needs to be made widely known in order that it can be fully exploited in solving problems in relevant areas. In this paper we review recent results in the area of learning automata, reexamine some of the theoretical questions that arise, and suggest potential areas where the available results may find application.

Brief Survey of Earlier Work

Historically, the first learning automata models were developed in mathematical psychology. Early work in this area has been well documented in the book by Bush and Mosteller [GB1]. More recent results can be found in Atkinson *et al.* [GA1]. A rigorous mathematical framework has been developed for the study of learning problems by Iosifescu and Theodorescu [GI1] as well as by Norman [GN1].

Tsetlin [DT1] introduced the concept of using deterministic automata operating in random environments as models of learning. A great deal of work in the Soviet Union and elsewhere has followed the trend set by his source paper. No attempt, however, has been made in this paper to review all these studies.

Varshavskii and Vorontsova [LV1] observed that the use of stochastic automata with updating of action probabilities could reduce the number of states in comparison with deterministic automata. This idea has proved to be very fruitful and has been exploited in a series of investigations, the results of which form the subject of this paper.

Fu and his associates [LF1]–[LF6] were among the first to introduce stochastic automata into the control literature. A variety of applications to parameter optimization, pattern recognition, and game theory were considered by this school. McLaren [LM1] explored the properties of linear updating schemes and suggested the concept of a "growing" automaton [LM2]. Chandrasekaran and Shen [LC1]–[LC3] made useful studies of nonlinear updating schemes, nonstationary environments, and games of automata. Tsypkin and Poznyak [LT1] attempted to unify the updating schemes by focusing attention on an inverse optimization problem. The present authors and their associates [LS1], [LS2], [LV3]–[LV10], [LN1], [LN2], [LL1]–[LL5] have studied the theory and applications of learning automata and also carried out simulation studies in the area.

The survey papers on learning control systems by Sklansky [GS1] and Fu [GF1] have devoted part of their attention to learning automata. The topic also finds a place in some books and collections of articles on learning systems [GM2], [GF2], [LF6]. The literature on the two-armed bandit problem is relevant in the present context but is not referred to in detail as the approach taken is rather different [LC5], [LW2]. References to other contributions will be made at appropriate points in the body of the paper.

Organization

This paper has been divided into nine sections. Following the introduction, the basic concepts and definitions of stochastic automata and random environments are given in Section II. The possible ways in which the behavior of learning automata can be judged are defined in Section III. Section IV deals with reinforcement schemes (or updating algorithms) and their properties and includes a discussion of convergence. Section V describes collective behavior of automata in terms of games between automata and multilevel structures of automata. Nonstationary environments are briefly considered in Section VI. Possible uses of learning automata in optimization and hypothesis testing form the subject matter of Section VII. A short description of the fields of application of learning automata is given in Section VIII. A comprehensive bibliography is provided in the Reference section and is divided into three subsections dealing with 1) general references in the literature pertinent to the topic considered, 2) some important papers on deterministic automata that provided the impetus for stochastic automata models, and 3) publications wholly devoted to learning automata.

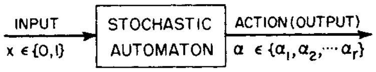

Fig. 1. Stochastic automaton.

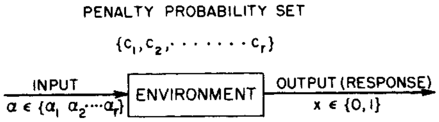

Fig. 2. Environment.

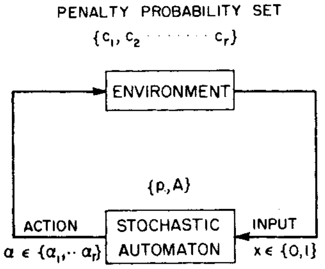

Fig. 3. Learning automaton.

II. Stochastic Automata and Random Environments

Stochastic Automaton

A stochastic automaton is a sextuple $\{x,\phi,\alpha,p,A,G\}$ where x is the input set, $\phi = \{\phi_1,\phi_2,\cdots,\phi_s\}$ is the set of internal states, $\alpha = \{\alpha_1,\alpha_2,\cdots,\alpha_r\}$ with $r \leq s$ is the output or action set, p is the state probability vector governing the choice of the state at each stage (i.e., at each stage n, $p(n) = (p_1(n),p_2(n),\cdots,p_s(n))^t$), A is an algorithm (also called an updating scheme or reinforcement scheme) which generates $p(n+1)$ from $p(n)$, and $G\colon \phi \rightarrow \alpha$ is the output function. G could be a stochastic function, but there is no loss of generality in assuming it to be deterministic [GP1]. In this paper G is taken to be deterministic and one-to-one (i.e., $r = s$, states and actions are regarded synonymous) and $s < \infty$. Fig. 1 shows a stochastic automaton with its inputs and actions.

It may be noted that the states of a stochastic automaton correspond to the states of a discrete-state discrete-parameter Markov process. Occasionally, it may be convenient to regard the $p_i(n)$ themselves as states of a continuous-state Markov process.

Environment

Only an environment (also called a medium) with random response characteristics is of interest in the problems considered. The environment (shown in Fig. 2) has inputs $\alpha(n) = \{\alpha_1,\cdots,\alpha_r\}$ and outputs (responses) belonging to a set x. Frequently the responses are binary $\{0,1\}$ with zero being called the nonpenalty response and one as the penalty response. The probability of emitting a particular output symbol (say, 1) depends on the input and is denoted by $c_i(i = 1,\cdots,r)$. The c_i are called the penalty probabilities. If the c_i do not depend on n, the environment is said to be stationary. Otherwise it is nonstationary. It is assumed that the c_i are unknown initially; the problem would be trivial if they are known *a priori.*

Learning Automaton (Stochastic Automaton in a Random Environment)

Fig. 3 represents a feedback connection of a stochastic automaton and an environment. The actions of the automaton in this case form the inputs to the environment. The responses of the environment in turn are the inputs to the automaton and influence the updating of the action probabilities. As these responses are random, the action probability vector $p(n)$ is also random.

In psychological learning experiments the organism under study is said to learn when it improves the probability of correct response as a result of interaction with its environment. Since the stochastic automaton being considered in this paper behaves in a similar fashion, it appears proper to refer to it as a learning automaton. Thus a learning automaton is a stochastic automaton that operates in a random environment and updates its action probabilities in accordance with the inputs received from the environment so as to improve its performance in some specified sense.

In the context of psychology, a learning automaton may be regarded as a model of the learning behavior of the organism under study and the environment as controlled by the experimenter. In an engineering application such as the control of a process, the controller corresponds to the learning automaton, while the rest of the system with all uncertainties constitutes the environment.

It is useful to note the distinction between several models based on the nature of the input to the learning automaton. If the input set is binary, e.g., $\{0,1\}$, the model is known as a P-model. On the other hand it is called a Q-model if the input set is a finite collection of distinct symbols as, for example, obtained by quantization and an S-model if the input set is an interval $[0,1]$. Each of these models appears appropriate in certain situations.

A remark on the terminology is relevant here. Following Tsetlin [DT1], deterministic automata operating in random environments have been proposed as models of learning behavior. Thus they are also contenders to the term "learning automata." However, in the view of the present authors the stochastic automaton with updating of action probabilities is a general model from which the deterministic automaton can be obtained as a special case having a 0–1-state transition matrix, and it appears reasonable to apply the term learning automaton to the more general model. In cases where it is felt necessary to emphasize the learning properties of a deterministic automaton one can use a qualifying term such as "deterministic learning automaton." It may also be noted that learning automata of this paper have been referred to as "variable-structure stochastic automata," in earlier literature [LV1].

III. Norms of Behavior of Learning Automata

The basic operation carried out by a learning automaton is the updating of the action probabilities on the basis of the responses of the environment. A natural question here is to examine whether the updating is done in such a manner as to result in a performance compatible with intuitive notions of learning.

One quantity useful in judging the behavior of a learning automaton is the average penalty received by the automaton. At a certain stage n, if the action α_i is selected with probability $p_i(n)$ the average penalty conditioned on $p(n)$ is

$$M(n) = E\{x(n) \mid p(n)\} = \sum_{i=1}^{r} p_i(n)c_i. \tag{1}$$

If no *a priori* information is available, and the actions are chosen with equal probability (i.e., at random), the value of the average penalty is denoted by M_0 and is given by

$$M_0 = \frac{c_1 + c_2 + \cdots + c_r}{r}. \tag{2}$$

The use of the term learning automaton can be justified if the average penalty is made less than M_0, at least asymptotically. Such a behavior is called expediency and is defined as follows [DT1], [LC1].

Definition 1: A learning automaton is called *expedient*[1] if

$$\lim_{n\to\infty} E[M(n)] < M_0. \tag{3}$$

When a learning automaton is expedient it only does better than one which chooses actions in a purely random manner. It would be desirable if the average penalty could be minimized by a proper selection of the actions. In such a case the learning automaton is called optimal. From (1) it can be seen that the minimum value of $M(n)$ is $\min_i \{c_i\}$.

Definition 2: A learning automaton is called *optimal* if

$$\lim_{n\to\infty} E[M(n)] = c_l \tag{4}$$

where

$$c_l = \min_i \{c_i\}.$$

Optimality implies that asymptotically the action associated with the minimum penalty probability is chosen with probability one. While optimality appears a very desirable property, certain conditions in a given situation may preclude its achievement. In such a case one would aim at a suboptimal performance. One such property is given by ε-optimality [LV4].

Definition 3: A learning automaton is called *ε-optimal* if

$$\lim_{n\to\infty} E[M(n)] < c_l + \varepsilon \tag{5}$$

can be obtained for any arbitrary $\varepsilon > 0$ by a suitable choice of the parameters of the reinforcement scheme. ε-optimality implies that the performance of the automaton can be made as close to the optimal as desired.

It is possible that the preceding properties hold only when the values of penalty probabilities c_i satisfy certain restrictions, for example, that they should lie in certain intervals. In such cases the properties are said to be conditional.

In practice, the penalty probabilities are often completely unknown, and it would be necessary to have desirable performance whatever be the values of c_i, that is, in all stationary random media. The performance would also be superior if the decrease of $E[M(n)]$ is monotonic. Both these requirements are considered in the following definition [LL3].

Definition 4: A learning automaton is said to be *absolutely expedient* if

$$E[M(n+1) \mid p(n)] < M(n) \tag{6}$$

for all n, all $p_k(n) \in (0,1)(k = 1,\cdots,r)$, and all possible values[2] of $c_i(i = 1,\cdots,r)$. Absolute expediency implies that $M(n)$ is a supermartingale and that $E[M(n)]$ is strictly monotonically decreasing with n in all stationary random environments. If $M(n) \leq M_0$ initially, absolute expediency implies expediency. It is thus a stronger requirement on the learning automaton. Furthermore, it can be shown that absolute expediency implies ε-optimality in all stationary random environments [LL4]. It is not at present known whether the reverse implication is true. However, every learning automaton presently known to be ε-optimal in all stationary media is also absolutely expedient. Hence ε-optimality and absolute expediency will be treated as synonymous in the sequel.

The definitions in this section have been given with reference to a P-model but can be applied with minor changes to Q- and S-models [LV3], [LV8], [LC1].

IV. Reinforcement Schemes

Having decided on the norms of behavior of learning automata we can now focus attention on the means of achieving the desired performance. It is evident from the description of the learning automaton that the crucial factor that affects the performance is the reinforcement scheme for the updating of the action probabilities. It thus becomes necessary to relate the structure of a reinforcement scheme and the performance of the automaton using the scheme.

In general terms a reinforcement scheme can be represented by

$$p(n+1) = T[p(n),\alpha(n),x(n)] \tag{7}$$

where T is an operator; $\alpha(n)$ and $x(n)$ represent the action of the automaton and the input to the automaton at instant n, respectively. One can classify the reinforcement schemes either on the basis of the property exhibited by a learning automaton using the scheme (as, for example, the automaton being expedient or optimal) or on the basis of the nature of the functions appearing in the scheme (as, for example, linear, nonlinear, or hybrid). If $p(n+1)$ is a linear function of the components of $p(n)$, the reinforcement scheme is said to be linear, otherwise it is nonlinear. Sometimes it is advantageous to update $p(n)$ according to different schemes depending on the intervals in which the value of $p(n)$ lies.

[1] Since $p_i(n)$, $\lim_{n\to\infty} p_i(n)$, and consequently $M(n)$ are, in general, random variables, the expectation operator is needed in the definition to represent the average penalty.

[2] It is usually assumed that the set $\{c_i\}$ has unique maximum and minimum elements.

In such a case the combined reinforcement scheme is called a hybrid scheme.

The basic idea behind any reinforcement scheme is rather simple. If the learning automaton selects an action α_i at instant n and a nonpenalty input occurs, the action probability $p_i(n)$ is increased, and all the other components of $p(n)$ are decreased. For a penalty input, $p_i(n)$ is decreased, and the other components are increased. These changes in $p_i(n)$ are known as reward and penalty, respectively. Occasionally the action probabilities may be retained at the previous values, in which case the status quo is known as "inaction."

In general, when the action at n is α_i

$$\begin{aligned} p_j(n+1) &= p_j(n) - f_j(p(n)), \qquad \text{for } x(n) = 0 \\ p_j(n+1) &= p_j(n) + g_j(p(n)), \qquad \text{for } x(n) = 1. \\ (j &\neq i) \end{aligned} \tag{8a}$$

The algorithm for $p_i(n + 1)$ is to be fixed so that $p_k(n + 1)$ $(k = 1,\cdots,r)$ add to unity. Thus

$$\begin{aligned} p_i(n+1) &= p_i(n) + \sum_{j \neq i} f_j(p(n)), \qquad \text{for } x(n) = 0 \\ p_i(n+1) &= p_i(n) - \sum_{j \neq i} g_j(p(n)), \qquad \text{for } x(n) = 1 \end{aligned} \tag{8b}$$

where the nonnegative[3] continuous functions $f_j(\cdot)$ and $g_j(\cdot)$ are such that $p_k(n + 1) \in (0,1)$, for all $k = 1,\cdots,r$ whenever every $p_k(n) \in (0,1)$. The latter requirement is necessary to prevent the automaton from getting trapped prematurely in an absorbing barrier.

Varshavskii and Vorontsova [LV1] were the first to suggest such reinforcement schemes for two-state automata and thus set the trend for later developments. They considered two schemes—one linear and the other nonlinear—in terms of updating of the state-transition probabilities. Fu, McLaren, and McMurtry [LF1], [LF2] simplified the procedure by considering updating of the total action probabilities as dealt with here.

Linear Schemes

The earliest known scheme can be obtained by setting

$$f_j(p) = ap_j \qquad g_j(p) = bp_j + b/r - 1, \qquad \text{for all } j = 1,\cdots,r \tag{9}$$

where $0 < a, b < 1$.[4] This is known as a linear reward–penalty (denoted L_{R-P}) scheme. Early studies of the scheme, principally dealing with the two-state case, were made by Bush and Mosteller [GB1] and Varshavskii and Vorontsova [LV1]. McLaren [LM1] made a detailed investigation of the multistate case, and this work was continued by Chandrasekaran and Shen [LC1] as well as by Viswanathan and Narendra [LV9]. Norman [LN4] established several results pertaining to the ergodic character of the scheme.

It is known that an automaton using the L_{R-P} scheme is expedient in all stationary random environments. Expressions for the rate of learning and the variance of the action probabilities are also available.

By setting

$$f_j(p) = ap_j \qquad g_j(p) \equiv 0, \qquad \text{for all } j \tag{10}$$

we get the linear reward–inaction (L_{R-I}) scheme. This scheme was considered first in mathematical psychology [GB1] but was later independently conceived and introduced into the engineering literature by Shapiro and Narendra [LS1], [LS2].

The characteristic of the scheme is that it ignores penalty inputs from the environment so that the action probabilities remain unchanged under these inputs. Because of this property a learning automaton using the scheme has been called a "benevolent automaton" by Tsypkin and Poznyak [LT1].

The L_{R-I} scheme was originally reported to be optimal in all stationary random environments, but it is now known that it is only ε-optimal [LV4], [LL4]. It is significant, however, that replacing the penalty by inaction in the L_{R-P} scheme totally changes the performance from expediency to ε-optimality.

Other possible combinations such as the linear reward–reward, penalty–penalty, and inaction–penalty schemes have been considered in [LV9], but these are, in general, inferior to the L_{R-I} and L_{R-P} schemes. The effect of varying the parameters a and b with n has also been studied in [LV9].

Nonlinear Schemes

As mentioned earlier, the first nonlinear scheme for a two-state automaton was proposed by Varshavskii and Vorontsova [LV1] in terms of transition probabilities. The total-probability version of the scheme corresponds to the choice

$$g_j(p) = f_j(p) = ap_j(1 - p_j), \qquad j = 1,2. \tag{11}$$

This scheme is ε-optimal in a restricted random environment satisfying either $c_1 < 1/2 < c_2$ or $c_2 < 1/2 < c_1$. Chandrasekaran and Shen [LC1] have studied nonlinear schemes with power-law nonlinearities. Several nonlinear schemes, which are ε-optimal in all stationary random environments, have been suggested by Viswanathan and Narendra [LV9] as well as by Lakshmivarahan and Thathachar [LL1], [LL3]. A simple scheme of this type for the two-state case is

$$f_j(p) = ap_j^2(1 - p_j) \qquad g_j(p) = bp_j^2(1 - p_j), \qquad j = 1,2 \tag{12}$$

where $0 < a < 4$, $0 < b < 1$.

A combination of linear and nonlinear terms often appears advantageous [LL3]. Extensive simulation results on a variety of schemes utilizing several possible combinations of reward, penalty, and inaction are available in [LV10]. A result that unifies most of the preceding reinforcement schemes has been reported in [LL3] and is given by the following.

[3] The nonnegativity condition need be imposed only if the "reward" character of $f_j(\cdot)$ and the "penalty" character of $g_j(\cdot)$ are to be preserved.

[4] $g_j(\cdot)$ for this scheme is not nonnegative for all values of p_j.

Theorem: A necessary and sufficient condition for the learning automaton using (8) to be absolutely expedient is

$$\frac{f_1(p)}{p_1} = \frac{f_2(p)}{p_2} = \cdots = \frac{f_r(p)}{p_r} = \lambda(p)$$

$$\frac{g_1(p)}{p_1} = \frac{g_2(p)}{p_2} = \cdots = \frac{g_r(p)}{p_r} = \mu(p) \qquad (13)$$

where $\lambda(\cdot)$ and $\mu(\cdot)$ are arbitrary continuous functions satisfying[5]

$$0 < \lambda(p) < 1$$

$$0 < \mu(p) < \min_j (p_j/1 - p_j) \qquad (14)$$

for all $j = 1,\cdots,r$ and all $p_j \in (0,1)$.

In simple terms, the theorem suggests that to obtain absolute expediency one type of updating should be made for the probability of the action selected by the automaton at the instant considered and a different type of updating for all the other action probabilities. No distinction should be made among the actions not selected by the automaton in the sense that the ratio $p_j(n + 1)/(p_j(n))$ should be the same for all these actions.

All the schemes known so far, which are ε-optimal in all stationary random environments, are also absolutely expedient; hence the functions appearing in these schemes satisfy the conditions of the theorem. The theorem also provides guidelines for the choice of new schemes.

Convergence of Action Probabilities

So far only the gross behavior of the automaton based on changes in the average penalty has been considered. It is of interest to probe deeper into the nature of asymptotic behavior of the action probability vector $p(n)$.

There are two distinct types of convergence associated with the reinforcement schemes. In one case the distribution functions of the sequence of action probabilities converge to a distribution function at all points of continuity of the latter function. This mode of convergence occurs typically in the case of expedient schemes such as the L_{R-P} scheme. A different mode of convergence occurs in the case of ε-optimal schemes (such as the L_{R-I} scheme). Here it can be proved using the martingale convergence theorem that the sequence of action probabilities converges to a limiting random variable with probability one. Thus a stronger mode of convergence is exhibited by ε-optimal schemes [LN4].

The difference between the two modes of convergence can be understood by the fact that expedient schemes (such as L_{R-P}) generate Markov processes that are ergodic but have no absorbing barrier, whereas ε-optimal schemes result in Markov processes with more than one absorbing barrier. Initial values of action probabilities do not affect the asymptotic behavior in the case of expedient schemes, whereas they play a crucial role in the case of ε-optimal schemes.

It has been shown that when the L_{R-P} scheme is used, $p(n)$ converges to a random variable with a continuous distribution in which the mean and variance can be computed. The variance can be made as small as desired by a proper choice of the parameters of the scheme [LM1], [LC1].

In ε-optimal schemes the action probability vector $p(n)$ converges to the set of absorbing states with probability one. As there are at least two such states and only one of the states is the desired one (i.e., the state associated with the minimum penalty probability) one can only say that $p(n)$ converges to the desired state with a positive probability. It is important to quantify this probability.

For simplicity, consider a two-state case. If $c_1 < c_2$, we require $p_1(n) \to 1$, and if $c_1 > c_2$, $p_1(n) \to 0$. When an ε-optimal scheme is used the only conclusion that can be drawn is $p_1(n) \to \{0,1\}$ with probability one, hence, the desirable event happens with a specific probability. Furthermore, this probability depends on the initial value $p_1(0)$. In order to gain confidence in the use of the schemes the probability of convergence to the desired state has to be determined as a function of the initial probability.

For fixing ideas, let us assume $c_1 < c_2$. It is necessary to find a function[6] $\gamma(p)$ defined by

$$\gamma(p) = \Pr\left[p_1(\infty) = 1 \mid p_1(0) = p\right].$$

The available results in the theory of stochastic stability are not of much use here as they treat, for the most part, the case of one absorbing state, whereas there are two such states here. Pioneering work on the present problem has been done by Norman [LN3], who has shown that $\gamma(p)$ satisfies a functional equation

$$U\gamma(p) = \gamma(p)$$

with suitable boundary conditions at $p = 0$ and $p = 1$, where the operator U is defined by

$$U\gamma(p) = E[\gamma(p_1(n + 1)) \mid p_1(n) = p].$$

However, this functional equation is extremely difficult to solve. Hence the next best thing that can be done is to establish upper and lower bounds on $\gamma(p)$. These can be computed by finding two functions $\psi_1(p)$ and $\psi_2(p)$ such that

$$U\psi_1(p) \geq \psi_1(p)$$

and

$$U\psi_2(p) \leq \psi_2(p)$$

for all $p \in [0,1]$ with appropriate boundary conditions. Satisfaction of these inequalities yields

$$\psi_2(p) \leq \gamma(p) \leq \psi_1(p).$$

The functions $\psi_1(\cdot)$ and $\psi_2(\cdot)$ are called subregular and superregular, respectively, and are comparatively easy to find as it involves only establishment of inequalities. Use of

[5] These conditions broadly arise because the $p_j(n)$ are probabilities and are required to lie in (0,1) and sum to unity. For a more precise statement see [LL3].

[6] p represents a scalar in the remainder of this subsection.

exponential functions appears particularly appropriate here. With

$$\psi_i(p) = \frac{\exp((x_i p) - 1)}{\exp(x_i - 1)}, \qquad i = 1,2$$

Norman [LN3] has established tight bounds for $\gamma(p)$ in the case of the L_{R-I} scheme with two states. The technique has been extended to cover nonlinear schemes and multistate cases [LL4].

Comments on Convergence Results: Recognizing the importance of the study of convergence of the action probabilities, several researchers have attempted to simplify the procedures involved. Some intuitively attractive ideas were employed extensively in this context, but on closer scrutiny they have been found to be incorrect. The mechanism of convergence now appears to be more complex than what was thought of earlier. Attention will be drawn to some of these fallacious arguments in the following.

1) In the case of a two-state automaton, suppose $p_1(n)$ satisfies

$$\begin{aligned} \Delta p_1(n) = E[p_1(n+1) - p_1(n) \mid p_1(n)] &> 0, && \text{for all } p_1(n) \in (0,1) \\ &= 0, && \text{for } p_1(n) = 0 \text{ and } 1. \end{aligned}$$

It follows immediately that $\lim_{n\to\infty} p_1(n) = \{0,1\}$ and that $E[p_1(n)]$ is strictly monotonically increasing. It has further been contended [LS1], [LL1] that since $E[p_1(n)]$ is bounded above by unity, $E[p_1(n)] \to 1$ as $n \to \infty$. This in turn implies that $p_1(n) \to 1$ with probability one. However, the conclusion that $E[p_1(n)] \to 1$ is not necessarily true even though $E[p_1(n)]$ is strictly monotonically increasing. Neither is the conclusion about convergence of $p_1(n)$ to unity with probability one.

2) A stability argument to be described has been widely used [LV1], [LC1] in gaining insight into the nature of convergence. Let p_{1i} $(i = 1,2,\cdots)$ be the roots (in the unit interval) of the martingale equation

$$\Delta p_1(n) = E[p_1(n+1) - p_1(n) \mid p_1(n)] = 0.$$

The roots satisfying

$$\left.\frac{d\Delta p_1}{dp_1}\right|_{p_1 = p_{1i}} < 0$$

are called stable roots and the rest unstable. It is then argued, on the basis of regarding Δp_1 as an increment in $p_1(n)$, that $p_1(n)$ converges to the set of stable roots with probability one. When $\Delta p_1(n)$ is sign definite the argument reduces to that in the previous comment 1).

On deeper probing, no justification of this argument has been found. It does not generally appear possible to prove convergence with probability one when there are roots of the martingale equation other than those corresponding to the absorbing states. Indeed what conclusions can be drawn in such a situation are at present unclear. The only situation, which can be handled presently (following Norman's approach outlined earlier) when absorbing states are present, is the case when all the roots of the martingale equation correspond to these absorbing states.

In view of the preceeding clarifications, some of the conclusions drawn in the earlier literature have to be modified. The nonlinear schemes regarded as optimal in restricted environments by Varshavskii and Vorontsova [LV1] as well as by Chandrasekaran and Shen [LC1] are now seen to be only ε-optimal. Similarly the L_{R-I} scheme of Shapiro and Narendra [LS1] and the nonlinear schemes given by Viswanathan and Narendra [LV9] Lakshmivarahan and Thathachar [LL1] are only ε-optimal in all stationary random environments. The nonlinear schemes in [LC1] where there are roots of the martingale equation other than those corresponding to the absorbing states have to be studied in greater detail in order to make definitive statements about their convergence. In connection with the L_{R-P} scheme, the asymptotic values of the state probabilities given in [LC1] actually refer to their expectations.

Some General Remarks on Reinforcement Schemes

1) All the schemes available in the literature so far are either expedient or ε-optimal. Vorontsova [LV2] stated sufficient conditions for a scheme to be optimal for a two-state automaton operating in continuous time. It is not known at present whether these conditions ensure optimality for the discrete-time automaton of our interest.

2) The question whether expedient or ε-optimal schemes are to be used in a particular situation is of considerable interest. While ε-optimal schemes show definite advantages in stationary random environments, the situation is not clear in the case of more general environments. A detailed comparison of various schemes has been made by Viswanathan and Narendra [LV6].

3) Only schemes suitable for the P-model have been discussed here. With appropriate modifications each of these schemes can be made suitable for Q- and S-models [LC1], [LV3], [LV8].

4) For the sake of simplicity, examples mostly of two-state schemes have been given. These can, however, be extended in a simple manner for use with multistate automata [LV9], [LL4].

5) Several linear and nonlinear schemes have been discussed in this section. Often improved rates of convergence can be obtained by a combination of linear and nonlinear terms [LL3], [LL4]. ε-optimal schemes are usually slow when the action probabilities are close to their final values. The convergence can be speeded up by using a hybrid scheme constructed from a judicious combination of an ε-optimal scheme and an expedient scheme [LV10].

6) In any particular scheme, the speed of convergence can be increased by changing the values of the parameters in the scheme. However, this is almost invariably accompanied by an increase in the probability of convergence to the undesired action(s) [LL4]. We meet the classical problem of speed versus accuracy. To score on both the counts, it appears that a careful selection of the nonlinear terms is necessary. Development of an analytical measure of the speed of convergence is necessary for further progress in this regard.

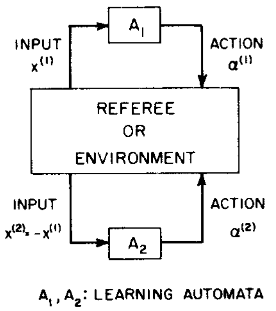

Fig. 4. Game between learning automata.

7) While the reinforcement scheme updates the action probabilities, a simultaneous estimation of the penalty probabilities of the environment using the Bayesian technique is proposed by Lakshmivarahan and Thathachar [LL2]. This leads to a fast estimate of the optimal action and also provides a confidence level on this estimate.

V. Interaction of Automata

The discussions made thus far have been centered around a single stochastic automaton interacting with a random environment. We shall now consider interactions between several automata. In particular, two types of interactions are of interest. In one case several automata are operating together in an environment either in a competitive or a cooperative manner so that we have a game situation. In the other case we consider a hierarchical system where there are various levels of automata, and there is interaction between different levels. Many interesting aspects of interaction, mostly with reference to deterministic automata, have been brought out in a series of papers by Varshavskii [DV5]–[DV8].

Games of Automata

Consider two automata operating in the same environment (Fig. 4). Each automaton selects its actions without any knowledge of the operation of the other automaton. For each pair of actions selected by the automata, the environment responds in a random manner with outcomes that form zero-sum inputs to the two automata. The action probabilities of the two automata are now independently updated according to suitable reinforcement schemes, and the procedure is repeated. The interest here is in the asymptotic behavior of the action probabilities and inputs.

This problem may be regarded as a game between the two automata. In this context several quantities can be redefined as follows. The event corresponding to the selection of a pair of actions by the automata is called a play. A game consists of a sequence of plays. Each action chosen by an automaton in a play is called a strategy. The environment is known as a referee, and the input received by each automaton corresponds to the payoff. Since the sum of the payoffs is zero, the game corresponds to a zero-sum two-person game. The asymptotic value of the expectation of the payoff to one of the players is called the value of the game.

In the classical theory of games developed by Von Neumann [GM1], [GL2], [GV1] it is assumed that the players are provided with complete information regarding the game, such as the number of players, their possible strategies, and the payoff matrix. The players can, therefore, compute their optimal strategies using this prior information. In the automata games being considered here such prior knowledge is not available, and the automata have to choose their strategies on-line. The payoff function is a random variable with an unknown distribution.

The concept of automata games was originally suggested by Krylov and Tsetlin [DK1], who considered competitive games of deterministic automata. These results were extended by the introduction of learning automata by Chandrasekaran and Shen [LC3], who used the L_{R-P} and nonlinear schemes. More recently Viswanathan and Narendra [LV7] discussed such games using ε-optimal schemes. It was demonstrated in [LV7] that when the payoff matrix has a unique saddle point, the value of the game coincides with that obtained in the deterministic situation, even though the characteristics of the game have to be learned as the game evolves. Since the automata operate entirely on the basis of their own strategies and the corresponding response of the referee, without any knowledge of the other automata participating in the game, this result also has implications in the context of decentralized control systems.

The preceding discussion pertains to competitive game problems, because each player, in trying to maximize his gain, is also attempting to minimize the gain of the other player. When the participants of a game cooperate with each other, we have a cooperative game. Stefanyuk and Vaisboard [DS1], [DV1], [DV2] considered cooperative games using deterministic automata. Viswanathan and Narendra [LV7] have shown that the value of the game can be made as close to the Von Neumann value as desired by using ε-optimal schemes when the payoff matrix has a unique saddle point.

The results available so far on learning automata games are few and limited. It appears that there is a broad scope for further study.

Multilevel Automata

A multilevel system of automata consists of several levels, each comprising many automata. Each action of an automaton at a certain level triggers automata at the next lower level, and thus the complete system has a tree structure. At each stage, decisions are made at various levels and communicated to lower levels in the hierarchy. The purpose of organizing such a multilevel system may be to achieve a performance that cannot be obtained using a single automaton or to overcome the high dimensionality of the decision space. The main problem in such multilevel systems is the coordination of activities at different levels or, in other words, to ensure convergence of the entire structure towards the desired behavior.

A two-level system has been proposed by Narendra and Viswanathan [LN1] for the optimization of performance of automata operating in a periodic random environment. It is assumed that the penalty probabilities $c_i(n)$ are periodic functions of n with an unknown period. The upper bound on the period is known.

The first level consists of one automaton in which the actions correspond to the various possible values of the period. The second level consists of a number of automata, one automaton corresponding to each of these first-level actions. Following the selection of a period $T(n)$ by the first level automaton, the first $T(n)$ automata in the second level are initiated to operate in the environment for one cycle. The average output of the environment in this cycle is used as the input to the first level to make the next selection of the value of the period. It has been shown through simulations [LN1] that the use of ε-optimal schemes leads to the convergence of the period to the true value.

VI. Nonstationary Environments

Most of the available work relates to the behavior of learning automata in stationary environments. The problem of behavior in nonstationary environments appears difficult, and only a few and specialized results are known [LL6], [LC4]. The interaction of automata in game situations and in hierarchical systems is one such special case. Some of the other known results will be described in this section.

Switched Random Environments

Tsetlin [DT1] considered a composite environment, which switches between a number of stationary environments in accordance with a Markov chain, and investigated the behavior of deterministic automata [DT2]. Varshavskii and Vorontsova [LV1] applied the learning automaton to the same problem and showed that expediency can be obtained for the entire range of possible values of the parameter of the switching transition probability matrix.

In a very limited situation where the Markov chain governing switching of the environment reaches its stationarity quickly in comparison with the time taken for the convergence of the reinforcement scheme, the problem essentially reduces to the operation in a stationary environment, hence, ε-optimal schemes would perform well.

Slowly-Varying Environments

When the penalty probabilities of the environment vary "slowly" in time, an ε-optimal scheme tends to lock on to a certain action, thereby losing its ability to change. As studied by Chandrasekaran and Shen [LC2], if the frequency of variation is sufficiently small, then the L_{R-P} scheme, which is expedient, seems to function satisfactorily. If prior information about the rate of variation is available, the performance of ε-optimal schemes can be improved by introducing reinitialization of action probabilities (that is, resetting them to be equal) at regular intervals of time.

Another possible approach appears to be to bound the action probabilities so that they fall short of attaining the absorbing barriers and are thus free to change according to changes in the environment. A comparison of the performances of the expedient linear scheme and ε-optimal schemes with the suggested modification is not available at the present time.

Periodic Environments

If it is known *a priori* that the penalty probabilities of the environment vary periodically in time with a common period T, the period T may be divided into N intervals. A system of N automata with a suitable arrangement can then be used so that one automaton is in operation at any instant of time and each automaton operates only once in every period. This is equivalent to each automaton operating in a stationary environment, so that over many cycles of operation the automata converge to the desired actions. In the case in which the environment is known to be periodic but the value of the period is unknown, a two-level automaton can be used as already described in Section V.

VII. Utilization of Automata

The learning automata discussed in earlier sections can be utilized to perform certain specific functions in a systems context. In particular, they may be used as optimizers or as decision-making devices and consequently can prove useful in a large class of practical problems.

Parameter Optimization

As remarked earlier many problems of adaptive control, pattern recognition, filtering, and identification can, under proper assumptions, be regarded as parameter optimization problems. It appears that a learning automaton can be fruitfully applied to solve such problems especially under noisy conditions when the *a priori* information is small. In fact, the possibility of using a stochastic automaton as a model for a learning controller provided the first motivation for studies in this area.

Given a system with only noisy measurements $g(\alpha,\omega)$ on the performance function $I(\alpha) = E\{g(\alpha,\omega)\}$, where α is an m-vector of parameters and ω is the measurement noise, the parameter optimization problem is to determine the optimal vector of parameters α_{opt} such that the system performance function is extremized. It is assumed that an analytical solution is not possible because of lack of sufficient information concerning the structure of the system and its performance function or because of mathematical intractability. The performance function $I(\alpha)$ may also be multimodal.

When this problem is tackled by gradient methods, both deterministic and stochastic, a search in the parameter space is carried out resulting in convergence to a local optimum. In the automaton approach the adaptive controller is the automaton in which the actions correspond to different values of α. The automaton updates the probabilities of the various parameter values based on the measurements of the performance function.

As observed earlier, the gradient methods are in a sense inhibited by the fact that at each instant a new value of the parameter is to be chosen close to the previous value. There is no similar restriction in the automaton approach, for each parameter value has a certain probability of being selected and only these probabilities are updated. Thus the learning automaton has the desired flexibility not to get locked on to a local optimum, and this difference makes automata methods particularly attractive for use in systems having multimodal performance criteria.

Several studies have been made on applying automata methods to parameter optimization. When a P-model automaton is used, one problem is to relate the performance measurement, which is usually a continuous real variable, to a binary-valued response of the environment. Fu and McLaren [LF2] avoided this by defining the penalty-strength (S-) model where the environmental response can take any value in [0,1]. McMurtry and Fu [LM3] applied a learning automaton using the L_{R-P} scheme to find the global minimum of a multimodal surface and showed that the automaton chooses the global minimum with the highest probability. The use of ε-optimal schemes such as the L_{R-I} scheme was advocated by Shapiro and Narendra [LS1], [LW1], who reported success on the difficult problem of handling a relatively flat performance surface along with the superposition of a high variance noise.

When the bounds on the performance measurements are known it is simple to normalize them to [0,1] and then use an S-model scheme. However, if these bounds are unknown, it is suggested in [LV3] that the current estimates of the bounds can be used for normalization, and it is further shown experimentally that ε-optimal convergence can still be achieved.

A problem of higher complexity was considered by Jarvis [LJ1], [LJ2] who studied, through simulation, the operation of a learning automaton using the L_{R-P} scheme as a global optimizer in a nonstationary environment. A pattern recognizer was used for sensing the changes in the environment.

The restriction that the set of parameter values considered must be finite is sometimes undesirable. To overcome this McLaren [LM2] has proposed a "growing automaton" where the number of actions of the automaton can grow countably to ∞. A comparison of several on-line learning algorithms, which include the growing automaton algorithm, was recently made by Saridis [GS2].

Problem of High Dimensionality: A basic problem associated with the use of automata methods in parameter optimization is that of high dimensionality. The problem is caused by the fact that the number of control actions of the automaton rapidly increases with the number of parameters and the "fineness" of the grid employed in the search. The speed of convergence of the updating schemes is to a large extent dependent upon the number of control actions, and thus when the parameter space contains a large number of points the convergence is too slow to be of any practical use. In case the parameter space is continuous, the automata method cannot be applied directly.

There are two methods of overcoming this problem of high dimensionality, as presented in the following discussion. Both the methods employ several interacting automata.

1) *Two-Level Controller:* The controller has a two-level structure here. The parameter space is divided into a finite number (say, r) of regions $v_1,v_2,\cdots,v_r$. There is one automaton at the first level having r actions each corresponding to one region, and this automaton acts as a supervisor governing the choice of one of the regions. When discretization of the parameter space is permissible, the second level has r automata, each automaton having control over one region. If, on the contrary, the parameter space is to be treated as continuous, the second level consists of a local search such as stochastic approximation. In this case it is necessary that the number of regions be large enough so that each region has at most one local extremum if the two-level system is to act as a global optimizer. Narendra and Viswanathan [LV5] demonstrated through simulation that the two-level system exhibits a faster convergence rate than a single automaton.

2) *Cooperative Games of Automata:* The results of cooperative games of automata can be naturally applied to the parameter optimization problem as mentioned in Section V. Consider the problem where there are m parameters each of which is discretized into r values. The parameter space now has r^m points. If r and m are large, the use of a single automaton controller would lead to a slow rate of convergence. Instead, if the controller consists of m automata each of which chooses the value of one parameter, then the number of probabilities to be updated at each stage would only be rm. The m automata used in this manner could be regarded as playing a cooperative game with the common object of extremizing the performance function. It follows from the results of Viswanathan and Narendra [LV5], [LV7] that if the performance function is unimodal, the use of ε-optimal schemes by the m-automaton controller leads to convergence of parameter values to the optimum with as high a probability as desired. Further research is needed to extend this approach to multimodal search problems.

Statistical Decision-Making

As the learning automaton selects the desirable action from the set of all actions on the basis of probabilistic inputs from the environment, one can regard the automaton as a device for making decisions under uncertainty. It can thus be expected to be used as a tool for solving statistical decision problems.

Many problems in control and communication can be posed as the fundamental problem of deciding between two hypotheses H_1 and H_2 on the basis of noisy observations $x(n)$. The conditional densities $p(x \mid H_1)$ and $p(x \mid H_2)$ are given, and the problem is to decide whether H_1 or H_2 is true so that the probabilities of making errors of the two kinds are less than prespecified values.

In order to apply the learning automaton approach to this situation it is necessary to make certain identifications. The two hypotheses H_1 and H_2 are made to correspond to the two actions of an automaton, and the observations $x(n)$ are regarded as the responses from the environment. Binary responses required for a P-model are obtained by using a threshold.

As no *a priori* information on the true hypothesis is available, the initial action probabilities are set at 0.5 each, and the automaton is allowed to operate according to an ε-optimal scheme. The hypothesis corresponding to the action in which the probability attains unity is taken as the true one. A design procedure for choosing the threshold and the parameters of the reinforcement scheme so as to satisfy any prespecified bounds on the error probabilities has been worked out by Lakshmivarahan and Thathachar [LL5]. Extension to multiple-hypothesis testing is also straightforward [LL4].

Testing Reinforcement Schemes

Attempts to develop new learning models in recent years have resulted in a variety of linear and nonlinear reinforcement schemes with adjustable parameters. It has consequently become increasingly difficult to compare two schemes for any given application. A novel approach suggested by Viswanathan and Narendra [LV6] is to set up a competitive game between two automata using the given schemes. A suitable game matrix with a unique saddle point is chosen, and the automata are allowed to play the game. The behavior of the two automata may then be used to determine the effects of parameters as well as the superiority of one of the schemes. Several such comparisons have been made in [LV6], and the superiority of ε-optimal schemes over expedient schemes in stationary environments has been demonstrated.

VIII. Applications

As mentioned in earlier sections, the theory of learning automata is still in its infancy, and the few significant results that are available were obtained only in recent years. Naturally, the application of these newly developed concepts to real-world problems has lagged somewhat, and only a few attempts have so far been reported.

An early application of the automaton approach was studied by Riordon [LR1], [LR2] in connection with the control of heat treatment regarded as a Markov process with unknown dynamics. The automaton acts as the adaptive controller in modeling the process as well as generating the appropriate control signals. Glorioso *et al.* [LG1], [LG2] have discussed adaptive routing in large communication networks using heuristic ideas of probabilistic iteration. Recently these ideas have been extended by Narendra, Tripathi, and Mason [LN5] to the telephone traffic routing problem for a real large-scale telephone network in British Columbia. Li and Fu [LL8] have applied a stochastic automaton to select features in agricultural data obtained by remote sensing. Several interesting applications to numerical analysis, random counters, and modeling appear to be contained in a recent Russian publication edited by Lorentz and Levin [LL7].

Other attempts have been made in the Soviet Union to apply automaton methods to a variety of situations, such as reliability [DG1], power regulation [DS2], switching circuits [DV3], queuing systems [DV4], and detection [DR1]. While most of these studies are concerned with deterministic automata models, there does not appear to be any reason why learning automata cannot be used in similar situations.

Learning automata provide a novel and computationally attractive mode of attacking a large class of problems involving uncertainties of a high order. As such they constitute an alternative approach to the well-known parameter optimization method using stochastic approximation. It is the opinion of the authors that a judicious combination of the two approaches will find increasing application in many practical problems in the future.

Acknowledgment

The authors would like to thank R. Viswanathan and S. Lakshmivarahan who worked with them on several aspects of learning automata.

References

General

[GA1] R. C. Atkinson, G. H. Bower, and E. J. Crothers, *An Introduction to Mathematical Learning Theory.* New York: Wiley, 1965.

[GB1] R. R. Bush and F. Mosteller, *Stochastic Models for Learning.* New York: Wiley, 1958.

[GF1] K. S. Fu, "Learning control systems—Review and outlook," *IEEE Trans. Automat. Contr.*, vol. 15, pp. 210–221, Apr. 1970.

[GF2] K. S. Fu, Ed., *Pattern Recognition and Machine Learning.* New York: Plenum, 1971.

[GI1] M. Iosifescu and R. Theodorescu, *Random Processes and Learning.* New York: Springer, 1969.

[GL1] R. D. Luce, *Individual Choice Behavior.* New York: Wiley, 1959.

[GL2] R. D. Luce and H. Raiffa, *Games and Decisions.* New York: Wiley, 1957.

[GM1] J. McKinsey, *Introduction to the Theory of Games.* New York: McGraw-Hill, 1952.

[GM2] J. M. Mendel and K. S. Fu, Eds., *Adaptive, Learning and Pattern Recognition Systems.* New York: Academic, 1970.

[GM3] J. M. Mendel, "Reinforcement learning models and their applications to control problems," *Learning Systems—A Symposium of the 1973 Joint Automatic Control Conf.* (ASME publication), pp. 3–18, 1973.

[GN1] M. F. Norman, *Markov Processes and Learning Models.* New York: Academic, 1972.

[GP1] A. Paz, *Introduction to Probabilistic Automata.* New York: Academic, 1971.

[GS1] J. Sklansky, "Learning systems for automatic control," *IEEE Trans. Automat. Contr.*, vol. AC-11, pp. 6–19, Jan. 1966.

[GS2] G. N. Saridis, "On-line learning control algorithms," *Learning Systems—A Symposium of the 1973 Joint Automatic Control Conf.* (ASME publication), pp. 19–52, 1973.

[GT1] Ya. Z. Tsypkin, *Adaptation and Learning in Automatic Systems.* New York: Academic, 1971.

[GV1] J. Von Neumann and O. Morgenstern, *Theory of Games and Economic Behavior.* Princeton, N.J.: Princeton Univ., 1947.

Deterministic Automata

[DG1] S. L. Ginzburg and M. L. Tsetlin, "Some examples of simulation of the collective behavior of automata," *Probl. Peredachi Informatsii*, vol. 1, no. 2, pp. 54–62, 1965.

[DK1] V. Yu Krylov and M. L. Tsetlin, "Games between automata," *Avtomat. Telemekh.*, vol. 24, pp. 975–987, July 1963.

[DR1] L. E. Radyuk and A. F. Terpugov, "Effectiveness of applying automata with linear tactic in signal detection systems," *Avtomat. Telemekh.*, vol. 32, pp. 99–107, Apr. 1971.

[DS1] V. L. Stefanyuk, "Example of a problem in the joint behavior of two automata," *Avtomat. Telemekh.*, vol. 24, pp. 781–784, June 1963.

[DS2] N. L. Stefanyuk and M. L. Tsetlin, "Power regulation in a group of radio stations," *Probl. Peredachi Informatsii*, vol. 3, no. 4, pp. 49–57, 1967.

[DT1] M. L. Tsetlin, "On the behavior of finite automata in random media," *Avtomat. Telemekh.*, vol. 22, pp. 1345–1354, Oct. 1961.

[DT2] H. Tsuji, M. Mizumoto, J. Toyoda, and K. Tanaka, "An automaton in the non-stationary random environment," *Inform. Sci.*, vol. 6, no. 2, pp. 123–142, Apr. 1973.

[DV1] E. M. Vaisbord, "Game of two automata with differing memory depths," *Avtomat. Telemekh.*, vol. 29, pp. 91–102, Mar. 1968.

[DV2] ——, "Game of many automata with various depths of memory," *Avtomat. Telemekh.*, vol. 29, pp. 52–58, Dec. 1968.

[DV3] V. I. Varshavskii, M. V. Meleshina, and M. L. Tsetlin, "Behavior of automata in periodic random media and the problem of synchronization in the presence of noise," *Probl. Peredachy Informatsii*, vol. 1, no. 1, pp. 65–71, 1965.

[DV4] ——, "Priority organization in queuing systems using a model of collective behavior," *Probl. Peredachi Informatsii*, vol. 4, no. 1, pp. 73–76, 1968.

[DV5] V. I. Varshavskii, "Collective behavior and control problems," in *Machine Intelligence*, *3*, D. Michie, Ed. Edinburgh: Edinburgh Univ., 1968.

[DV6] ——, "The organization of interaction in collectives of automata," *Machine Intelligence*, *4*, B. Meltzer and D. Michie, Eds. Edinburgh: Edinburgh Univ., 1969.

[DV7] ——, "Some effects in the collective behavior of automata," *Machine Intelligence*, *7*, B. Meltzer and D. Michie, Eds. Edinburgh: Edinburgh Univ., 1972.

[DV8] ——, "Automata games and control problems," presented at the 5th World Congr. IFAC, Paris, France, June 12–17, 1972.

Learning Automata

[LC1] 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.

[LC2] ——, "Adaptation of stochastic automata in nonstationary environments," *Proc. NEC*, vol. 23, pp. 39–44, 1967.

[LC3] ——, "Stochastic automata games," *IEEE Trans. Syst. Sci. Cybern.*, vol. SSC-5, pp. 145–149, Apr. 1969.

[LC4] L. D. Cockrell and K. S. Fu, "On search techniques in switching environment," *Proc. 9th Symposium Adaptive Processes*, Austin, Tex., Dec. 1970.

[LC5] T. M. Cover and M. E. Hellman, "The two-armed bandit problem with time-invariant finite memory," *IEEE Trans. Inform. Theory*, vol. IT-16, pp. 185–195, Mar. 1970.

[LF1] K. S. Fu and G. J. McMurtry, "A study of stochastic automata as models of adaptive and learning controllers," Purdue Univ., Lafayette, Ind., Tech. Rep., TR-EE 65-8, 1965.

[LF2] K. S. Fu and R. W. McLaren, "An application of stochastic automata to the synthesis of learning systems," Purdue Univ., Lafayette, Ind., Tech. Rep. TR-EE 65-17, 1965.

[LF3] K. S. Fu, "Stochastic automata, stochastic languages and pattern recognition," *J. Cybern.*, vol. 1, no. 3, pp. 31–49, July–Sept. 1971.

[LF4] K. S. Fu and T. J. Li, "Formulation of learning automata and automata games," *Inform. Sci.*, vol. 1, no. 3, pp. 237–256, July 1969.

[LF5] ——, "On stochastic automata and languages," *Inform. Sci.*, vol. 1, no. 4, pp. 403–419, Oct. 1969.

[LF6] K. S. Fu, "Stochastic automata as models of learning systems," Computer and Information Sciences II, J. T. Tou, Ed. New York: Academic, 1967.

[LG1] R. M. Glorioso, G. R. Grueneich, and J. C. Dunn, "Self organization and adaptive routing for communication networks," *1969 EASCON Rec.*, pp. 243–250.

[LG2] R. M. Glorioso and G. R. Grueneich, "A training algorithm for systems described by stochastic transition matrices," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-1, pp. 86–87, Jan. 1971.

[LG3] B. R. Gaines, "Stochastic computing systems," in *Advances in Information Sciences 2*. New York: Plenum, 1969, pp. 37–172.

[LJ1] R. A. Jarvis, "Adaptive global search in a time-invariant environment using a probabilistic automaton," *Proc. IREE*, Australia, pp. 210–226, 1969.

[LJ2] ——, "Adaptive global search in a time-variant environment using a probabilistic automaton with pattern recognition supervision," *IEEE Trans. Syst. Sci. Cybern.*, vol. SSC-6, pp. 209–217, July 1970.

[LL1] S. Lakshmivarahan and M. A. L. Thathachar, "Optimal nonlinear reinforcement schemes for stochastic automata," *Inform. Sci.*, vol. 4, pp. 121–128, 1972.

[LL2] ——, "Bayesian learning and reinforcement schemes for stochastic automata," *Proc. Int. Conf. Cybernetics and Society*, Washington, D.C., Oct. 1972.

[LL3] ——, "Absolutely expedient learning algorithms for stochastic automata," *IEEE Trans. Syst. Man, Cybern.*, vol. SMC-3, pp. 281–286, May 1973.

[LL4] S. Lakshmivarahan, "Learning algorithms for stochastic automata," Ph.D. dissertation, Dep. Elec. Eng., Indian Institute of Science, Bangalore, India, Jan. 1973.

[LL5] S. Lakshmivarahan and M. A. L. Thathachar, "Hypothesis testing using variable-structure stochastic automata," in *1973 IEEE Decision and Control Conf.*, *Preprints*, San Diego, Calif.

[LL6] G. Langholtz, "Behavior of automata in a nonstationary environment," *Electron. Lett.*, vol. 7, no. 12, pp. 348–349, June 17, 1971.

[LL7] A. A. Lorentz and V. I. Lenin, Eds., *Probabilistic Automata and Their Applications* (in Russian). Riga, Latvia, USSR: Zinatne, 1971.

[LL8] T. J. Li and K. S. Fu, "Automata games, stochastic automata and formal languages," Purdue Univ., Lafayette, Ind., Tech. Rep. TR-EE69-1, Jan. 1969.

[LM1] R. W. McLaren, "A stochastic automaton model for synthesis of learning systems," *IEEE Trans. Syst. Sci. Cybern.*, vol. SSC-2, pp. 109–114, Dec. 1966.

[LM2] ——, "A stochastic automaton model for a class of learning controllers," in 1967 *Joint Automatic Control Conf.*, *Preprints*.

[LM3] G. J. McMurtry and K. S. Fu, "A variable-structure automaton used as a multimodal search technique," *IEEE Trans. Automat. Contr.*, vol. AC-11, pp. 379–387, July 1966.

[LM4] L. G. Mason, "Self optimizing allocation systems," Ph.D. dissertation, University of Saskatchewan, Saskatoon, Sask., Canada, 1972.

[LM5] ——, "An optimal learning algorithm for *S*-model environments," *IEEE Trans. Automat. Contr.*, vol. AC-18, pp. 493–496, Oct. 1973.

[LN1] K. S. Narendra and R. Viswanathan, "A two-level system of stochastic automata for periodic random environments," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-2, pp. 285–289, Apr. 1972.

[LN2] ——, "Learning models using stochastic automata," in *Proc. 1972 Int. Conf. Cybernetics and Society*, Washington, D.C., Oct. 9–12.

[LN3] M. F. Norman, "On linear models with two absorbing barriers," *J. Math. Psychol.*, vol. 5, pp. 225–241, 1968.

[LN4] ——, "Some convergence theorems for stochastic learning models with distance diminishing operators," *J. Math. Psychol.*, vol. 5, pp. 61–101, 1968.

[LN5] K. S. Narendra, S. S. Tripathi, and L. G. Mason, "Application of learning automata to telephone traffic routing problems," Becton Center, Yale Univ., New Haven, Conn., Tech. Rep. CT-60, Jan. 1974.

[LR1] J. S. Riordon, "Optimal feedback characteristics from stochastic automaton models," *IEEE Trans. Automat. Contr.*, vol. AC-14, pp. 89–92, Feb. 1969.

[LR2] ——, "An adaptive automaton controller for discrete-time Markov processes," *Automatica* (IFAC Journal), vol. 5, no. 6, pp. 721–730, Nov. 1969.

[LS1] 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, Oct. 1969.

[LS2] I. J. Shapiro, "The use of stochastic automata in adaptive control," Ph.D. dissertation, Dep. Eng. and Applied Sci., Yale Univ., New Haven, Conn., 1969.

[LS3] Y. Sawaragi and N. Baba, "A consideration on the learning behavior of variable structure stochastic automata," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-3, pp. 644–647, Nov. 1973.

[LT1] Ya. Z. Tsypkin and A. S. Poznyak, "Finite learning automata," *Eng. Cybern.*, vol. 10, pp. 478–490, May–June 1972.

[LV1] V. I. Varshavskii and I. P. Vorontsova, "On the behavior of stochastic automata with variable structure," *Avtomat. Telemekh.*, vol. 24, pp. 353–360, Mar. 1963.

[LV2] I. P. Vorontsova, "Algorithms for changing stochastic automata transition probabilities," *Probl. Peredachi Informatsii*, vol. 1, no. 3, pp. 122–126, 1965.

[LV3] R. Viswanathan and K. S. Narendra, "Stochastic automata models with applications to learning systems," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-3, Jan. 1973.

[LV4] ——, "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.

[LV5] R. Viswanathan, "Learning automata: models and applications," Ph.D. dissertation, Dep. Eng. and Applied Sci., Yale Univ., New Haven, Conn., 1972.

[LV6] R. Viswanathan and K. S. Narendra, "Comparison of expedient and optimal reinforcement schemes for learning systems," *J. Cybern.*, vol. 2, no. 1, pp. 21–37, 1972.

[LV7] ——, "Competitive and cooperative games of variable-structure stochastic automata," in *Joint Automatic Control Conf.*, *Preprints*, Aug. 1972; also available as Becton Center, Yale Univ., New Haven, Conn., Tech. Rep. CT-44, Nov. 1971.

[LV8] ——, "Application of stochastic automata models to learning systems with multimodal performance criteria," Becton Center, Yale Univ., New Haven, Conn., Tech. Rep. CT-40, June 1971.

[LV9] ——, "Expedient and optimal variable-structure stochastic automata," Dunham Lab., Yale Univ., New Haven, Conn., Tech. Rep. CT-31, Apr. 1970.

[LV10] ——, "Simulation studies of stochastic automata models, Part I—Reinforcement schemes," Becton Center, Yale Univ., New Haven, Conn., Tech. Rep. CT-45, Dec. 1971.

[LW1] I. H. Witten, "Comments on 'Use of stochastic automata for parameter self-optimization with multimodal performance criteria,' " *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-2, Apr. 1972.

[LW2] ——, "Finite-time performance of some two-armed bandit controllers," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-3, pp. 194–197, Mar. 1973.