Decentralized Learning in Finite Markov Chains

RICHARD M. WHEELER, JR., MEMBER, IEEE, AND KUMPATI S. NARENDRA, FELLOW, IEEE

Abstract—**The principal contribution of this paper is a new result on the decentralized control of finite Markov chains with unknown transition probabilities and rewards. One decentralized decision maker is associated with each state in which two or more actions (decisions) are available. Each decision maker uses a simple learning scheme, requiring minimal information, to update its action choice. It is shown that, if updating is done in sufficiently small steps, the group will converge to the policy that maximizes the long-term expected reward per step. The analysis is based on learning in sequential stochastic games and on certain properties, derived in this paper, of ergodic Markov chains. A new result on convergence in identical payoff games with a unique equilibrium point is also presented.**

I. Introduction

Decentralized decision making requiring limited information is a highly desirable feature of large complex systems. In some cases, these systems can be described only by imprecise models and yet decentralized methods still may be necessary. This paper presents one such method, applicable to an important class of stochastic systems, finite state Markov decision processes. It is shown that decentralized decision makers, located at the states of a finite Markov chain, can effectively use simple learning schemes to update their decisions. The principal result derived (Theorem 3) is that, without prior knowledge of transition probabilities or rewards, the collection of controllers will converge to the set of actions that maximizes the long-term expected reward per unit time obtained by the system.

Finite state Markov decision processes (controlled Markov chains) arise when state transitions generate rewards or costs which depend on actions taken in some or all states. Both finite time and infinite time problems with various performance criteria have been posed and means of determining the optimal policy via dynamic programming methods are well known [1]–[3].

However, several important factors have limited the applicability of this type of control. First, the computation becomes burdensome when the number of states, although finite, is very large. Second, often the information about the model required for a centralized approach such as dynamic programming is not available. Specifically, transition probabilities and corresponding rewards associated with various actions may be unknown at the time control is begun or may change during system operation. This leads to an adaptive problem in which, typically, parameters are estimated and, using a separation principle, the subsequent estimates are used to update control actions. This problem formulation has attracted recent interest and several results on convergence of the parameter estimates and control policy are now available [4]–[6]. Finally, uncertainty may be present in the form of incomplete or noisy state observation. Versions of this problem have been solved (e.g., [7], [8]), but imperfect state observation coupled with the adaptive problem remains a difficult open problem.

The learning approach presented in this paper addresses primarily the adaptive problem, although much of the computational difficulty mentioned above is avoided as well. The model differs from that of other adaptive approaches in that no explicit dependence of the system behavior on an unknown parameter is assumed, and hence no explicit estimation is required. In contrast, the setting is one of myopic local agents, one located at each state, who are unaware of the surrounding world. There is no knowledge that other agents exist or indeed that the world is an N-state Markov chain whose transition probabilities and corresponding rewards depend on actions chosen. Each local agent, in attempting to improve its performance, simply chooses an action, waits for a response, and then updates its action. Significantly, no prior computation must be done to solve for the optimal policy assuming the parameters were known, a nontrivial problem in itself when N is large.

There is no explicit synchronization of decision makers in this formulation. The algorithm at any state is updated only at those instants when the process returns to that state. The updating is done via a simple learning scheme which uses the cumulative reward obtained from a given action normalized by the total elapsed time under that action as its environmental response. The intriguing and powerful result is that individuals operating in nearly total ignorance of their surroundings can implicitly coordinate themselves to lead to optimal group behavior. This result is based on a new result on learning in N-player identical payoff games, presented in Section III, and exploits a special property of ergodic Markov chains, discussed in Section IV. Some simple examples are included in Section V to demonstrate the convergence behavior of the learning schemes.

II. Control of Markov Chains

The control of finite Markov chains for which transition probabilities and rewards are known is a well-known problem (e.g., [9], [1]) and can be stated as follows. Let $\Phi = \{\phi_1, \phi_2, \cdots, \phi_N\}$ be the state space of a finite Markov chain $\{x_n\}_{n\geq 0}$, and $\boldsymbol{\alpha}^i = \{\alpha_1^i, \alpha_2^i, \cdots, \alpha_{r_i}^i\}$ be the finite set of actions (also called controls or decisions) available in state ϕ_i. The transition probabilities $t_j^i(k)$ and rewards $r_j^i(k)$ depend on the starting state ϕ_i, the ending state ϕ_j, and the action $\alpha_k^i (k = 1, \cdots, r_i)$ used in ϕ_i. The goal is to choose the set of actions, or policy, $\boldsymbol{\alpha} = \{\alpha_{i_1}^1, \alpha_{i_2}^2, \cdots, \alpha_{i_N}^N\} \in \mathcal{A} = \boldsymbol{\alpha}^1 \otimes \boldsymbol{\alpha}^2 \otimes \cdots \otimes \boldsymbol{\alpha}^N$ that maximizes

$$J(\boldsymbol{\alpha}) \triangleq \lim_{n\to\infty} \frac{1}{n} E\left[\sum_{t=0}^{n-1} r(x(t), x(t+1), \boldsymbol{\alpha})\right] \tag{1}$$

where $r(x(t), x(t+1), \boldsymbol{\alpha})$ is the reward generated by a transition from $x(t)$ to $x(t+1)$ using policy $\boldsymbol{\alpha}$. The set of policies is limited in this formulation to stationary nonrandomized policies. Hence, the best strategy in any state is a pure strategy and is independent of the time at which the state is occupied. Under the following assumption, it can be shown that the optimal $\boldsymbol{\alpha}$ in fact belongs to this set of policies [2].

Manuscript received December 17, 1984; revised August 5, 1985 and December 23, 1985. Paper recommended by Past Associate Editor, H. K. Khalil. This work was supported by the Schlumberger Corporation and the Systems Research Division of the ITT Advanced Technology Center.

R. M. Wheeler, Jr. was with the Center for Systems Science, Department of Electrical Engineering, Yale University, New Haven, CT 06520. He is now with Sandia National Laboratories, Livermore, CA 94550.

K. S. Narendra is with the Center for Systems Science, Department of Electrical Engineering, Yale University, New Haven, CT 06520.

IEEE Log Number 8608339.

Assumption 1: The Markov chain corresponding to each policy $\boldsymbol{\alpha}$ is ergodic.

Thus, there are no transient states and a limiting distribution $\pi(\boldsymbol{\alpha}) = (\pi_1(\boldsymbol{\alpha}), \pi_2(\boldsymbol{\alpha}), \cdots, \pi_N(\boldsymbol{\alpha}))$ exists, with each $\pi_i(\boldsymbol{\alpha}) > 0$, which is independent of the initial state. Under Assumption 1, the expected reward per step $J(\boldsymbol{\alpha})$ defined in (1) can also be written in terms of the limiting (stationary) probabilities $\pi(\boldsymbol{\alpha})$ as

$$J(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \pi_i(\boldsymbol{\alpha}) \sum_{j=1}^{N} t_j^i(\boldsymbol{\alpha}) r_j^i(\boldsymbol{\alpha}). \tag{2}$$

The above expression is used to derive an interesting and essential relationship between changes in $\boldsymbol{\alpha}$ and relative values of $J(\boldsymbol{\alpha})$ (Theorem 2 of Section IV).

A dynamic programming solution that optimizes (1) [or equivalently (2)] when full information is available was given in [1], but the computational cost of the scheme increases dramatically with increasing N. Other methods have been proposed which provide an easily computable bound on the difference in performance between a given policy and the optimal policy [10], [11]. A two-layer scheme with control algorithms at the two layers operating on different time scales has also been suggested as a means of dealing with high dimensionality [12]. However, the "large-scale" problem introduced by many states remains a major practical limitation of many theoretical results. In this respect, decentralization is highly desirable and thus plays a central role in the problem formulation in this paper.

An entirely new set of difficulties arises when the transition probabilities or rewards, assumed known above, are not available. In this case, information about the unknown system must be learned as control action is taking place. Thus, the optimal policy cannot be found off-line even if the computational problem can be overcome. This adaptive problem can be stated in terms of an unknown parameter θ upon which the $t_j^i(k, \theta)$ depend (e.g., [4]–[6]). All the uncertainty is embodied in θ and the rewards (or costs) are assumed known. The general approach is to estimate θ at each instant and then use the action that would be optimal if the estimate were the true value. Probabilistic learning schemes have also been applied to this version of the adaptive problem in [13], [14], although in [14] a discounted reward criterion is used.

The problem as formulated in this paper is strikingly different from that just described. Specifically, no knowledge of either $t_j^i(k)$ or $r_j^i(k)$ is assumed, other than that the $r_j^i(k)$ are normalized to lie in the interval [0, 1]. Since control is effected in a direct method, no explicit parameter estimation is needed. Most important, the control scheme is implemented in a decentralized fashion. While this problem formulation was treated in [15], no globally optimal adaptive scheme was given.

The preceding discussion is summarized below.

Statement of the Problem: Let the controlled Markov chain $\{x_n\}_{n \geq 0}$ occupy state ϕ_i at time n, where $\phi_i \in \Phi$, the set of N states. States in which at least two actions are available are distinguished as *action states.* Clearly, if Φ^* is the set of all action states ($|\Phi^*| = N^* \leq N$), then $\Phi^* \subseteq \Phi$. If $\phi_i \in \Phi^*$, decision maker $A_i (i = 1, \cdots, N^*)$ chooses an action $\alpha_k^i \in \boldsymbol{\alpha}^i$. This action consists of the probability vector $t^i(k)$ governing the state transition from ϕ_i and a corresponding reward vector $r^i(k)$. The chain then moves from ϕ_i to ϕ_j and a reward $r_j^i(k)$ is generated. Both $t_j^i(k)$ and $r_j^i(k)$ are unknown. If $\phi_i \notin \Phi^*$, there is only one transition vector t^i and one reward vector r^i governing the chain.

Problem: Find a decentralized control scheme for each A_i so that asymptotically the collection of decision makers $\{A_i\}$ will evolve to policy $\boldsymbol{\alpha}^* \in \mathfrak{A}$ where $J(\boldsymbol{\alpha}^*) = \max_{\boldsymbol{\alpha}} \{J(\boldsymbol{\alpha})\}$ and $J(\boldsymbol{\alpha})$ is given by (1) or (2).

Comments:

i) In contrast to the methods referred to above, a key element of the desired control scheme is that no prior computation must be done to solve for the optimal policy for each value of an unknown parameter.

ii) It will be shown below that only a minimum of information exchange among the decentralized decision makers is required. Even the number of states N is not used by any local decision maker.

iii) As in any stochastic convergence problem, asymptotic behavior must be qualified. The scheme proposed below converges to $\boldsymbol{\alpha}^*$ with probability arbitrarily close to one (w.p. $1 - \epsilon$), where ϵ can be made as small as desired by adjusting a parameter in the algorithms.

The adaptive scheme introduced in this paper makes use of learning automata and the analysis is based on convergence properties of such automata in stochastic games. While a vast literature exists on learning schemes that achieve various types of performance, only the basic features necessary to derive the key result on learning in games (Theorem 1) are presented in Section III. It is the combination of properties of Markov chains and the behavior of automata in games that enables the total solution to the problem stated above.

III. The Learning Approach

Learning models have been studied extensively by psychologists and systems theorists from both modeling and control viewpoints [16]–[20]. A particularly simple model for sequential decision making in unknown random environments is the learning automaton (see [21], [22] for extensive bibliographies). While many variants of this model have been proposed and studied, the underlying idea can be simply stated. The automaton has a finite set of actions and sequentially updates a probability distribution over the action set, based on the environmental response to the stimulus of a particular action chosen. In this manner, the automaton attempts to improve its behavior in some sense, i.e., learn. Typically, it is assumed that the response is a random variable whose distribution, while stationary, is different for each automaton action.

Single Automaton Model: Let $\boldsymbol{\alpha} = \{\alpha_1, \alpha_2, \cdots, \alpha_r\}$ be the automaton action set and $\boldsymbol{\beta}$ be the environment response set. Each element of $\boldsymbol{\beta}$ is quantified and for convenience normalized to lie in the interval [0, 1], where 1 corresponds to the best response and 0 to the worst. If at time n the action, response, and action probability vector are denoted by $\alpha(n) \in \boldsymbol{\alpha}$, $\beta(n) \in \boldsymbol{\beta}$, and $p(n) = \{p_1(n), p_2(n), \cdots, p_r(n)\}$, respectively, then the manner in which $p(n)$ is updated is governed by the learning algorithm T where $p(n + 1) = T[p(n), \alpha(n), \beta(n)]$. The specification of T constitutes the design of the automaton. Although many nonlinear schemes have been studied, the essential behavioral properties of any automaton can be exhibited by the following linear algorithm with various parameter values.

If $\alpha(n) = \alpha_i$, then

$$p_i(n+1) = p_i(n) + a\beta(n)[1 - p_i(n)] - b[1 - \beta(n)]p_i(n)$$

$$p_j(n+1) = p_j(n) - a\beta(n)p_j(n) + b(1 - \beta(n))\left[\frac{1}{r-1} - p_j(n)\right] \qquad j \neq i \tag{3}$$

where $0 < a < 1$ and $0 \leq b < 1$ are constants called the reward and penalty parameters, respectively. It is assumed that all initial probabilities $p_i(0)$ lie in the open interval (0, 1). If $b = a$ the scheme is called linear reward-penalty ($L_{R\text{-}P}$); if $b \ll a$ it is called linear reward-ϵ-penalty ($L_{R\text{-}\epsilon\text{-}P}$); and if $b = 0$ it is called linear reward-inaction ($L_{R\text{-}I}$). While any of these schemes could be used as decentralized controllers, in the remainder of the paper attention will be restricted to the $L_{R\text{-}I}$ scheme.

Convergence Properties: Convergence of the $L_{R\text{-}I}$ version of (3) is demonstrated by observing that it gives rise to a Markov process with stationary transition probabilities. The vector $p(n)$ defined by (3) is a random vector. If the distributions of the response $\beta(n)$ to the various actions α_i are independent of time, then the probability of $p(n + 1)$ is determined completely by

$p(n)$. Further, if for any n the unit simplex S_r is defined as

$$S_r \triangleq \left\{ p \,|\, 0 \le p_i \le 1,\ \sum_{i=1}^{r} p_i = 1 \right\} \tag{4}$$

then the learning scheme represents the mapping $T\colon S_r \to S_r$. Hence, $\{p(n)\}_{n \ge 0}$ is a discrete parameter Markov process defined on the state space S_r with a stationary transition function.

An important feature of the linear algorithm (3) for any a and b is the distance-diminishing property, which makes the resulting Markov process compact. The mapping T defined on S_r is said to be distance diminishing on S_r (i.e., T is a stochastic contraction) if, for any points p^1 and p^2 belonging to S_r,

$$\sup_{p^1 \ne p^2} \frac{d(T[p^1, \alpha, \beta],\ T[p^2, \alpha, \beta])}{d(p^1, p^2)} < 1 \tag{5}$$

where $d(x^1, x^2)$ is the Euclidean metric. The convergence of the $L_{R\text{-}I}$ scheme then amounts to the asymptotic behavior of compact Markov processes. While much is known about such processes, the following important convergence result is sufficient in this paper.

If $\beta(n)$ is viewed as a measure of relative success, with $\beta(n) = 1$ equal to the maximum success, then the expected success is defined as

$$M(n) \triangleq E[\beta(n)|p(n)] = \sum_{i=1}^{r} E[\beta(n)|\alpha_i]p_i(n). \tag{6}$$

Defining $d_0 = \max_i \{E[\beta(n)|\alpha_i]\}$, Norman [18] has shown the following.

Theorem: For any $\epsilon > 0$ and any $p(0)$ in the open simplex S_r, there exists an $0 < a^* < 1$ such that for $b = 0$ and any $a < a^*$ in (3)

$$\lim_{n \to \infty} E[M(n)] > d_0 - \epsilon.$$

Proof: The proof is given in [18], for the case in which $\beta(n)$ is a binary random variable ($\boldsymbol{\beta} = \{0, 1\}$).

The proof of this result, termed *ϵ-optimality,* for the general $\beta(n)$ is similar [23]. All automata convergence results stated or derived in this paper are in terms of ϵ-optimality.

Automata Games and Decentralized Decision Making: Much of the original inquiry into automata attempted to model decentralized control in biological systems [20]. By contrast, decentralization is not an issue in the model as discussed above. However, it becomes the principal issue when many automata are interconnected as in the Markov chain control problem considered in this paper. The close connection between such problems and abstract games was recently identified in [24]. It is this realization that motivates the discussion of automata games below, including the important Theorem 1.

An *automata game* involves N automata $A_i (i = 1, \cdots, N)$, each with an action (strategy) set $\boldsymbol{\alpha}^i = \{\alpha_1^i, \alpha_2^i, \cdots, \alpha_{r_i}^i\}$, interacting through a stationary random environment. At each instant n, each A_i selects one action according to its current probability distribution $p^i(n) = (p_1^i, p_2^i, \cdots p_{r_i}^i)$. The joint action (or *play*) $\boldsymbol{\alpha}(n) = \boldsymbol{\alpha} = \{\alpha_{i_1}^1, \alpha_{i_2}^2, \cdots, \alpha_{i_N}^N\}$, chosen with probability $p(\boldsymbol{\alpha}) = \Pi_{j=1}^N p_{i_j}^j(n)$, determines the distribution of the random responses $\beta^i(n)$ received. Stationarity comes from the fact that the response distributions are fixed over time. Each automaton has access only to its own response. In contrast to usual game-theoretic formulation, no player is aware of other players, the play chosen at any instant or any of the response distributions.

In the Markov chain problem, the automaton at each state probabilistically chooses an action, waits for a response, and then updates its action probabilities. The response received is the only information available to A_i and it depends upon actions taken by other automata when the chain is in other states. In this respect, the decentralized control of Markov chains bears a strong similarity to an automata game, even though the automata are not synchronized in the former case as they are in the latter.

The performance of automaton A_i in the game is judged, as in the single automaton case, by the asymptotic behavior of $E[M^i(n)]$ where $M^i(n) \triangleq E[\beta^i(n)|p(n) = (p^1(n), p^2(n), \cdots, p^N(n))]$. Defining $d^i(\boldsymbol{\alpha}) = E[\beta^i(n)|\boldsymbol{\alpha}]$, then

$$M^i(n) = \sum_{\text{all } \alpha} p(\boldsymbol{\alpha}) d^i(\boldsymbol{\alpha}). \tag{7}$$

However, the notion of ϵ-optimality is not in general well defined for the game, since individual and group rationality are usually inconsistent. A notable exception is the two-person zero-sum game for which the value solution is well established as rational behavior. Although learning algorithms of the type (3) were designed as clever devices in simple environments, a powerful result is that they demonstrate group rationality (converge to the value) in the zero-sum game as well [25], [26].

Rational behavior, and therefore ϵ-optimality, in games with identical payoffs ($d^i(\boldsymbol{\alpha}) = d(\boldsymbol{\alpha}),\ \forall\ i$) is also well defined and amounts to choosing actions resulting in the optimum payoff, by definition the same for individual and group. Clearly the control of Markov chains falls into this class of game, since there is a single performance index to be optimized. The concept of an equilibrium point in strategy space is important in the identical payoff game, as in the zero-sum game.

Definition: Let $\alpha_{j_i}^i$ be the j_i^{th} strategy of player i and $M^i(\alpha_{j_1}^1, \alpha_{j_2}^2, \cdots, \alpha_{j_N}^N)$ be player i's payoff. Then $\boldsymbol{\alpha}^\circ = (\alpha_{j_1}^1{}^\circ, \alpha_{j_2}^2{}^\circ, \cdots, \alpha_{j_N}^N{}^\circ)$ is an *equilibrium* if $M^i(\alpha^\circ) = \max_{j_i} M^i(\alpha_{j_1}^1{}^\circ, \alpha_{j_2}^2{}^\circ, \cdots, \alpha_{j_i}^i, \cdots, \alpha_{j_N}^N{}^\circ)$ for all i.

Nash [27] proved that all finite games in strategic form have at least one pure or mixed strategy equilibrium. In the two-player identical payoff game the payoff structure can be represented by a game matrix. Any element of the matrix which is simultaneously maximum in its row and column is a pure strategy equilibrium. This property generalizes in a straightforward manner to the N-player game.

The following result on the convergence of learning schemes in identical payoff games plays a central role in the proof of convergence of the decentralized automata in the controlled Markov chain presented in Section IV. It is stated below with an outline of the proof, but the complete proof is contained in the Appendix.

Theorem 1: Let Γ be an identical payoff game among N players $A_i (i = 1, \cdots, N)$, each with r_i actions, and let all A_i use identical $L_{R\text{-}I}$ learning schemes. Let Γ have a *unique* pure strategy equilibrium $\boldsymbol{\alpha}^* = (\alpha_1^1, \alpha_1^2, \cdots, \alpha_1^N)$ with corresponding expected success $d(\boldsymbol{\alpha}^*)$ (each A_i uses its first action).

If the above conditions are satisfied, then for any $\epsilon > 0$, there exists an $0 < a^* < 1$ such that for $b = 0$ and any $a < a^*$ in (3)

$$\lim_{n \to \infty} E[M(n)] > d(\boldsymbol{\alpha}^*) - \epsilon.$$

Outline of Proof: The proof of Theorem 1 is presented in two stages. In the first stage, we consider an ordinary differential equation (ODE)

$$\dot{f} = W[f]$$

where f ranges over an $r_1 \times r_2 \times \cdots \times r_N$ simplex. The zeros of $W(\cdot)$ (defined below) are the stationary points of the ODE. It is demonstrated that under the conditions specified by Theorem 1, all but one of the stationary points are unstable. Hence, any solution of the ODE with initial conditions in the interior of the simplex converges to the stable solution denoted by f^*.

In the second stage, the distance-diminishing property (5) is shown to hold in the game, implying that the overall Markov process is compact. It follows that $p(n)$ converges to the set of ergodic kernels (absorbing states) of the process with probability 1 [18]. Next, the stochastic difference equation

$$\Delta p(n) = E[p(n+1) - p(n) | p(n)] \triangleq aW[p(n)]$$

is considered, where $p(n) = (p^1(n), p^2(n), \cdots, p^N(n))$. Since $E[p(n)|p(0) = p] - f(n) = k_1 a$ and $E[(p(n) - f(n))^2] = k_2 a$ for all n, it follows that $\lim_{n \to \infty} Pr\{p(n) = f^*\} > 1 - \epsilon$ can be realized by the proper choice of the step size a. From this the statement of Theorem 1 is easily obtained.

Comments:

i) Theorem 1 can be viewed in terms of stochastic hill-climbing. With a small a, many steps can be taken without moving very far from the current position, and hence averaging can occur, resulting in expected progress in the upward direction. For any number of equilibria it has been shown that $\Delta M(n) \geq 0$ for any initial probabilities and all n.

ii) Special cases of the conditions of Theorem 1 have been treated. Initial formulations of the automata game appear in [28]–[30] and a partial analysis of the two-player game is found in [31]. If the restrictive property of dominance holds, then a result for a limited class of nonstationary environments [32] is applicable. This has been exploited in the two-player game where each player has two actions to show ϵ-optimality [15].

IV. Decentralized Learning as a Control Strategy for Markov Chains

This section consists of three parts. In Section IV-A an updating scheme is proposed for the decentralized controllers in a controlled Markov chain. In Section IV-B a property of Markov chains is derived relating changes in policies $\boldsymbol{\alpha}$ to an ordering of the corresponding $J(\boldsymbol{\alpha})$. This result (Theorem 2) is of particular importance to decentralized control. Finally, using Theorems 1 and 2, it is shown in Section IV-C that the collection of updating schemes used in the problem of Section II is globally ϵ-optimal (Theorem 3).

A. The Automaton Updating Procedure

The decentralized learning control of a Markov chain as proposed here involves one learning automaton for each action state and a coordinator to perform simple administrative tasks. Each automaton operates on its own local time scale $n_i = \{0, 1, 2, \cdots\}$ while the chain, along with the coordinator, operates on the global time scale $n = \{0, 1, 2, \cdots\}$. If $x(n) = \phi_i \in \Phi^*$, then A_i, activated by the coordinator, chooses an action α^i_k according to its current action probability vector $p^i(n_i)$. Otherwise, no control is taken and the chain is governed by a fixed transition vector and reward vector.

A central feature of the control scheme is that A_i is not informed of the one-step reward $r^i_j(k)$ resulting from its action. In fact, A_i receives no information whatsoever about the effect of its action or about the activity of the chain until the process returns to ϕ_i. At that time $n_i + 1$, A_i receives two pieces of data from the coordinator: 1) the cumulative reward generated by the process up to time n, and 2) the current global time n. From these. A_i computes the incremental reward $\Delta\rho^i_k(n_i)$ generated since $n_i(\alpha^i(n_i) = \alpha^i_k)$ and the corresponding elapsed global time $\Delta\eta^i_k(n_i)$. These increments are added to their current cumulative totals $\rho^i_k(n_i)$ and $\eta^i_k(n_i)$, resulting in new totals $\rho^i_k(n_i + 1)$ and $\eta^i_k(n_i + 1)$. The environment response (input to A_i) is then taken to be

$$\beta^i(n_i+1) = \frac{\rho^i_k(n_i+1)}{\eta^i_k(n_i+1)}. \tag{8}$$

Since the $r^i_j(k)$ are normalized to lie in [0, 1], clearly $\beta^i(n_i)$ also lies in [0, 1].

The Algorithm: Each A_i is assumed to use the updating scheme (3), with $\beta(n)$ defined by (8). The modified scheme is denoted by $T1$.

The above procedure can be summarized as follows.

i) If at time n, $x(n) = \phi_i \in \Phi^*$, only A_i updates its action probabilities.

ii) If $x(n) \in \Phi - \Phi^*$ no control is taken, but the reward generated and one instant of global time are added to their respective current totals by the coordinator.

iii) In the intervening time between two visits to any ϕ_i, no knowledge of the sequences of states visited is provided to A_i. Only the current values of total reward and n are needed and these are incremented at each n whether or not $x(n) \in \Phi^*$.

Note that the role of the coordinator, while essential to the scheme, is that of bookkeeper, not decision maker. All estimation and control functions are performed by the decentralized automata.

B. A Property of Ergodic Finite Markov Chains

Let an N-state Markov chain have N_A action states $\phi^*_i \in \Phi^*$ with action set $\boldsymbol{\alpha}^i = \{\alpha^i_1, \alpha^i_2, \cdots, \alpha^i_{r_i}\}$, $r_i \geq 2$, in each state. Associate one decision maker A_i with each ϕ^*_i. Then $\Gamma = \{N_A, \mathfrak{A}, J\}$ denotes a finite identical payoff game among $\{A_i\}$ in which the play $\boldsymbol{\alpha} \in \mathfrak{A} = \boldsymbol{\alpha}^1 \otimes \boldsymbol{\alpha}^2 \otimes \cdots \otimes \boldsymbol{\alpha}^{N_A}$ results in the payoff $J(\boldsymbol{\alpha})$, where $J(\boldsymbol{\alpha})$ is given by (2).

Theorem 2: Γ has a unique equilibrium.

Proof: For convenience, assume that a dummy decision maker with only one action is also associated with each nonaction state. In the corresponding N-player game Γ', a play (policy) has N rather than N_A components. However, since the action sets in $N - N_A$ states are degenerate, any play $\boldsymbol{\alpha}$ in Γ is equivalent to a play in Γ' in which the N_A action state decision makers use $\boldsymbol{\alpha}$. We will prove the theorem for Γ'.

The proof is related to that given by Howard [1] for the convergence of the policy iteration method.[1]

Assume that a play α is a nonoptimal equilibrium point (EP) of Γ'. In Howard's terminology, the gain $J(\alpha)$ and the relative values $v^i(\alpha)$ associated with α are found as the solution to

$$J(\alpha) = q^i(\alpha) + \sum_{j=1}^{N} t^i_j(\alpha) v^j(\alpha) - v^i(\alpha), \qquad i = 1, \cdots, N \tag{9}$$

where $q^i(\alpha) = \Sigma^N_{j=1} t^i_j(\alpha) r^i_j(\alpha)$ and $v^N(\alpha)$ is set to zero arbitrarily to guarantee a unique solution. Using policy iteration, a better play than α can be found. Assume that state i is one of the states in which the better play differs from α. Consider the play β which differs from α only in state i. The component of β used in state i is found as that k which maximizes the following "test quantity"

$$\tau_i(k, \alpha) = q^i(k) + \sum_{j=1}^{N} t^i_j(k) v^j(\alpha) - v^i(\alpha) \tag{10}$$

over all $k = 1, \cdots, r_i$. The test quantity depends on α since the values $v^j(\alpha)$ in (10) are kept as the ones computed from the original play α. A useful property of the test quantities is that for any play β, $J(\beta) = \Sigma^N_{j=1} \pi_j(\beta)\tau_j(\beta, \alpha)$, where $\pi_j(\beta)$ is the steady state probability for state j under play β and $\tau_j(\beta, \alpha)$ is the test quantity formed in state j by using play β in state j but relative values associated with play α [1]. Assumption 1 assures the existence of the $\pi_j(\beta)$ for all j and β. From the maximization of (10),

$$\tau_i(\beta, \alpha) > \tau_i(\alpha, \alpha)$$

$$\tau_j(\beta, \alpha) = \tau_j(\alpha, \alpha), \qquad j \neq i.$$

[1] The authors are indebted to an anonymous reviewer who suggested this approach instead of a lengthier proof given in an earlier version of the paper.

Noting from (9) and (10) that $\tau_j(\alpha, \alpha) = J(\alpha)$ $(j = 1, \cdots, N)$, it follows that

$$J(\beta) = \sum_{j=1}^{N} \pi_j(\beta)\tau_j(\beta, \alpha) > \sum_{j=1}^{N} \pi_j(\beta)\tau_j(\alpha, \alpha) = J(\alpha).$$

Since β is superior to α and differs from α only in one state, α cannot be an EP, leaving the optimal play in Γ' as the unique EP.

Comment: The uniqueness of an equilibrium in Γ is an intriguing property. In any controlled Markov chain satisfying Assumption 1, the second best policy differs from the optimal policy only in the action chosen in exactly one state. In general, the kth best policy can differ from the optimal policy in at most the actions chosen in $k - 1$ states.

C. Convergence of the Decentralized Controllers

In the above discussion, the controlled Markov chain was represented as a game Γ, which was shown to have a unique equilibrium. If Γ were an automata game, then players using learning scheme (3) would be ϵ-optimal. However, a payoff in Γ is obtained asymptotically by using a fixed policy. Since each decision maker uses the updating procedure $T1$, Γ can only be viewed as a limiting game $\Gamma = \lim_{n \to \infty} \Gamma(n)$. The elements of $\Gamma(n)$, $s(\boldsymbol{\alpha}, n) \triangleq E[\beta(n)|\boldsymbol{\alpha}(n) = \boldsymbol{\alpha}]$, depend on n and thus the Markov chain updating is not the same as the updating in an automata game. However, from Assumption 1 it follows that $\lim_{n \to \infty} s(\boldsymbol{\alpha}, n) = J(\boldsymbol{\alpha})$. For a sufficiently large n the ordering among the $s(\boldsymbol{\alpha}, n)$ will be identical to that among the $J(\boldsymbol{\alpha})$ in Γ. Therefore, it is sufficient to analyze the automata game Γ.

The fact that $T1$ results in asynchronous updating, while the updating in an automata game is synchronous, is not important as long as the ratio of updating frequencies of any two controllers does not tend to zero with increasing n. It follows from Assumption 1 that this cannot happen.

Defining $J(n) = \Sigma_{\text{all } \alpha}\, p(n)s(\alpha, n)$, the principal result of the paper can be stated as the following theorem.

Theorem 3: Let an automaton A_i using learning scheme $T1$ be associated with each of the N_A action states, each having r_i actions, of an N state Markov chain. If Assumption 1 is satisfied, then for any $\epsilon > 0$, there exists an $0 < a^* < 1$ such that for $b = 0$ and any $a < a^*$ in $T1$

$$\lim_{n \to \infty} E[J(n)] > J(\boldsymbol{\alpha}^*) - \epsilon.$$

Proof: The proof follows immediately from Theorems 1 and 2.

Comment: In the above discussion, it is assumed, for ease of exposition only, that all of the automata use the same reward parameter a. This may not be desirable in decentralized control problems and is by no means necessary for decentralized learning results, as has been shown elsewhere (see, e.g., [26]).

V. Simulations

Two decentralized learning experiments, each with two controllers A_1 and A_2, are described below. The first deals with an identical payoff game as discussed in Section III (summarized in Theorem 1). The second treats the Markov chain control problem (summarized in Theorem 3). The mean learning curve is denoted by $E[p_1(n)] = (E[p_1^1(n)], E[p_1^2(n)])$, while $p_1(n) = (p_1^1(n), p_1^2(n))$ denotes a typical sample path. In each case, the mean learning curve was computed by averaging over 30 sample paths.

Experiment 1—Two-Player Identical Payoff Game: The game played by A_1 and A_2 is represented by the matrix

$$\Gamma_1 = \begin{bmatrix} 0.8 & 0.6 \\ 0.1 & 0.4 \end{bmatrix},$$

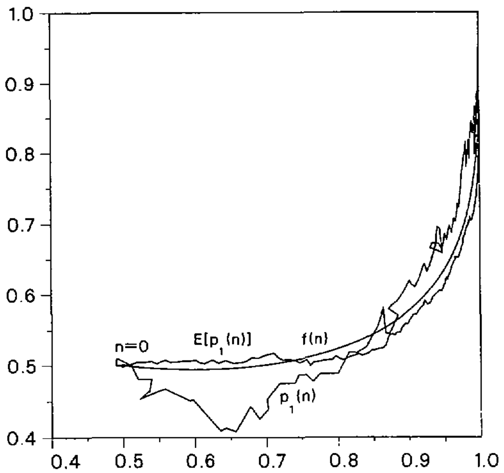


Fig. 1. Mean learning curve $E[p_1(n)]$, typical sample path $p_1(n)$, and approximate solution of associated deterministic differential equation $f(n)$ for Experiment 1.

where the actions of A_1 and A_2 correspond to the rows and columns, respectively, of Γ_1. The environment model is assumed to be binary, and hence $\beta(n)$ can only assume the values 1(success) or 0(failure). Each element of Γ_1 is the expected success resulting from the corresponding action pair.

In the experiment, A_1 and A_2 both used algorithm (3) with $a = 0.02$ and $b = 0$. The results are shown in Fig. 1, in which the horizontal and vertical axes represent the probabilities of the first actions $p_1^1(n)$ and $p_1^2(n)$ of A_1 and A_2, respectively. All 30 sample paths converged to the optimal action pair and the mean curve converged to (0.95, 0.95) in approximately 1000 steps. The smooth curve $f(n)$ is the approximation, valid to within $0(a^2)$, to the solution of the deterministic differential equation

$$\dot{f}(t) = W[f(t)], \; f(0) = p$$

(see the proof of Theorem 1 in the Appendix). It is clear that the difference between $E[p_1(n)]$ and $f(n)$ is small for any n.

Experiment 2—Controlled Markov Chain, $N = 2$, $N_A = 2$: In this example, both states are action states. Two actions are available in each state. The transition probabilities and rewards are summarized below

		Action	
		1	2
State	ϕ_1	$t^1(1) = (0.1\ 0.9)$	$t^1(2) = (0.9\ 0.1)$
		$r^1(1) = (0.2\ 0.2)$	$r^1(2) = (0.3\ 0.3)$
	ϕ_2	$t^2(1) = (0.1\ 0.9)$	$t^2(2) = (0.9\ 0.1)$
		$r^2(1) = (0.7\ 0.7)$	$r^2(2) = (0.8\ 0.8)$

Each controller uses the updating procedure $T1$ described in Section IV, with $a = 0.1$ and $b = 0$. It follows that the corresponding asymptotic game is

$$\Gamma_2 = \begin{bmatrix} 0.65 & 0.5 \\ 0.5 & 0.35 \end{bmatrix},$$

and hence choosing the first action in each state is the optimal policy. In the simulation, all of the 30 sample paths converged to the optimal policy. The mean curve converged to (0.95, 0.95) in approximately 900 steps of global time. Note that in each state, a higher immediate reward is obtained if the second action is chosen. However, the choice of $\beta(n)$ used in updating scheme $T1$ prohibits the biasing of the control policy in favor of short-term rewards.

VI. Conclusion and Comments

Two basic problems are addressed in this paper: i) decentralized decision making (and corresponding sequential games) and ii) adaptation with minimal prior information. At a gross level, both problems amount to the stability of algorithms and many papers have been devoted to showing convergence of particular algorithms for either problem. However, the approach presented here uniquely combines the structure of finite Markov chains with the flexibility of learning theory, resulting in decentralized adaptation which is globally ϵ-optimal. Several observations about the learning approach are summarized below.

i) The approach as developed here requires no global model to be shared by the various controllers. Very little information is either assumed *a priori* or exchanged during operation. Also, no precomputation of the optimal policy is required.

ii) The criterion of long-term expected reward per step is only one criterion that can be specified. It appears possible to achieve ϵ-optimality with respect to other reward criteria by suitably modifying the environment response $\beta(n)$.

iii) $\beta(n)$ in the updating algorithm is computed as the ratio of two numbers, each tending to infinity. Any practical implementation of the scheme would require a modified method of computing $\beta(n)$.

iv) Another implementation issue, speed of convergence, is particularly important in tracking changes in the transition probabilities or rewards that might occur. While the $L_{R\text{-}I}$ algorithm analyzed here is quite slow for small values of a, other related learning schemes have been suggested which exhibit faster convergence (e.g., [33]).

v) As mentioned in the Introduction, the method requires perfect state observation in order to activate the proper controller. The problem of noisy state observation raises difficult theoretical questions.

vi) Since the use of the learning approach is independent of the Markov chain structure exploited here, the application of decentralized learning algorithms to other collective decision-making problems appears promising. A preliminary inquiry into such problems is made in [24].

Many of the above comments suggest possible extensions or modifications to the problem statement or the learning scheme. However, the basic idea presented in this paper constitutes a new and important result in decentralized learning—simple and intuitive local schemes operating in an unknown environment can lead to globally optimal behavior.

Appendix

Proof of Theorem 1

As mentioned in Section III, the proof of Theorem 1 is given in two stages (A and B below). The following notation is used throughout this Appendix.

Let Γ be the N-player game with player A_i having r_i actions. Let A_i's action probability vector at stage n be $p^i(n) \in S_{r_i}$ where S_{r_i} is the unit simplex in R^{r_i}. The joint action probability vector at n is then $p(n) = (p^1(n), p^2(n), \cdots, p^N(n)) \in S = S_{r_1} \times S_{r_2} \times \cdots \times S_{r_N}$. Let $e^i_{i_j} \triangleq (0, 0, \cdots, 0, 1, 0, \cdots, 0) \in S_{r_j}$ where the i_jth element is 1. Then $V = \{e = (e^1_{i_1}, e^2_{i_2}, \cdots, e^N_{i_N})\}$ is the set of all vertices of S. Note that there is a unique $\boldsymbol{\alpha}$ corresponding to each e.

Define $\Delta p(n) \triangleq E[p(n+1) - p(n) | p(n) = p] = aW[p]$. Using the $L_{R\text{-}I}$ algorithm (3) (and omitting the argument n), it is readily shown that

$$W^i_j(p) = p^i_j \sum_{k \neq j}^{r_i} \sum_{\boldsymbol{\alpha} \setminus \alpha^i_k} p(\boldsymbol{\alpha} \setminus \alpha^i_k)[d(\boldsymbol{\alpha} \setminus \alpha^i_j) - d(\boldsymbol{\alpha} \setminus \alpha^i_k)] \quad \text{(A.1)}$$

where $p(\boldsymbol{\alpha} \setminus \alpha^i_k)$ and $d(\boldsymbol{\alpha} \setminus \alpha^i_k)$ denote the following. Let $\boldsymbol{\alpha} = (\alpha^1_{i_1}, \alpha^2_{i_2}, \cdots, \alpha^i_{i_i}, \cdots, \alpha^N_{i_N})$.

i) If $p(\boldsymbol{\alpha}) = \Pi^N_{j=1} p^j_{i_j}$, then $p(\boldsymbol{\alpha} \setminus \alpha^i_k) = \Pi^N_{j \neq i} p^j_{i_j} p^i_k$ and is the probability of the play

$$\boldsymbol{\alpha} \setminus \alpha^i_k = (\alpha^1_{i_1}, \alpha^2_{i_2}, \cdots, \alpha^i_k, \cdots, \alpha^N_{i_N}).$$

ii) If the expected success for play $\boldsymbol{\alpha}$ is $d(\boldsymbol{\alpha})$, then $d(\boldsymbol{\alpha} \setminus \alpha^i_k)$ is the expected success for the play $\boldsymbol{\alpha} \setminus \alpha^i_k$.

A. Stability of the Stationary Points of the ODE $f = W(f)$

The stationary points of the ODE $f = W(f)$ are the zeros of $W(f)$. Every $e \in V$ is a zero of $W(f)$ and it is shown below that only one element of V, $e^* = (e^1_1, e^2_1, \cdots, e^N_1)$, is stable. Hence, any solution of the ODE with an initial condition in the interior of S converges to the one stable state e^*. The following property is needed.

Property 1 (Unique Equilibrium Property of Γ*):* The definition of an equilibrium in a game, given in Section III, can be restated as follows. The play $\boldsymbol{\alpha}$ is an equilibrium of Γ if $d(\boldsymbol{\alpha}) > d(\boldsymbol{\alpha} \setminus \alpha^i_k)$ $\forall$ i and $\forall$ $k \neq i_i$. In Theorem 1, $\alpha^* = (\alpha^1_1, \alpha^2_1, \cdots, \alpha^N_1)$ is assumed to be the unique equilibrium.

Following stability arguments used in [34] and [31], we show that the point e is stable if

$$\left.\frac{dW^i_j}{dp^i_j}\right|_{p=e} < 0 \;\forall\; i \text{ and } j. \quad \text{(A.2)}$$

Note that the derivative in (A.2) can be expressed as

$$\frac{dW^i_j}{dp^i_j} = \frac{\partial W^i_j}{\partial p^i_j} + \sum_{m \neq j}^{r_i} \frac{\partial W^i_j}{\partial p^i_m} \cdot \frac{dp^i_m}{dp^i_j}. \quad \text{(A.3)}$$

The second factor in the summation in (A.3) can be shown to be negative by defining arbitrary constants λ^i_k, satisfying

$$p^i_k = \lambda^i_k p^i_2, \;\lambda^i_2 = 1;\; \lambda^i_k \geq 0, \qquad k = 3, 4, \cdots, r_i. \quad \text{(A.4)}$$

From (A.4) and $\Sigma^{r_i}_{k=1} p^i_k = 1$ $\forall$ i, it follows that

$$\frac{dp^i_k}{dp^i_j} = \frac{-\lambda^i_k}{\sum\limits_{m \neq j}^{r_i} \lambda^i_m} < 0 \;\forall\; i \text{ and } \forall\; j \neq k. \quad \text{(A.5)}$$

The expression (A.3) is now evaluated for all e by considering three separate cases.

Case i) Stability of $e^* = (e^1_1, e^2_1, \cdots, e^N_1)$: First consider dW^i_1/dp^i_1. From (A.1) it is found that

$$\left.\frac{\partial W^i_1}{\partial p^i_1}\right|_{e^*} = \sum_{k \neq 1}^{r_i} \sum_{\boldsymbol{\alpha} \setminus \alpha^i_k} p(\boldsymbol{\alpha} \setminus \alpha^i_k)[d(\boldsymbol{\alpha} \setminus \alpha^i_1) - d(\boldsymbol{\alpha} \setminus \alpha^i_k)]\Big|_{e^*} = 0 \quad \text{(A.6)}$$

and, using Property 1,

$$\left.\frac{\partial W^i_1}{\partial p^i_m}\right|_{e^*} = \sum_{\boldsymbol{\alpha} - \{i\}} p(\boldsymbol{\alpha} - \{i\})[d(\boldsymbol{\alpha} \setminus \alpha^i_1) - d(\boldsymbol{\alpha} \setminus \alpha^i_m)]\Big|_{e^*}$$

$$= d(\alpha^*) - d(\alpha^* \setminus \alpha^i_m) > 0 \quad \text{(A.7)}$$

where $\boldsymbol{\alpha} - \{i\}$ is the play chosen by the $N-1$ players, excluding A_i. Combining (A.3) with (A.5)–(A.7), it follows that $dW^i_1/dp^i_1|_{e^*} < 0$. A similar argument reveals that $dW^i_j/dp^i_j|_{e^*} = d(\alpha^* \setminus \alpha^i_j) - d(\alpha^*) < 0$, $j \neq 1$. Hence, e^* is stable.

Case ii): Stability of $e = (e^1_1, e^2_1, \cdots, e^i_k, \cdots, e^N_1)$ $k \neq 1$: Using the same argument as above, it follows that

$$\left.\frac{dW^i_1}{dp^i_1}\right|_e = d(\alpha^*) - d(\boldsymbol{\alpha} \setminus \alpha^i_k) > 0,$$

a sufficient condition for e of this type to be unstable.

Case iii): Stability of $e = (e^1_{i_1}, e^2_{i_2}, \cdots, e^N_{i_N})$ $i_k \neq 1$ $\forall$ k: Choosing any element W^i_j, it can be shown as above that

$$\left.\frac{dW^i_j}{dp^i_j}\right|_e = d(\boldsymbol{\alpha}) - d(\boldsymbol{\alpha} \setminus \alpha^i_{i_k}) = \gamma^i_k.$$

Stability requires that $\gamma^i_k < 0$ for all i and k, but this is impossible from Property 1. Hence, all other e's are unstable and e^* is the only stable element of V.

The fact that $p(n)$ consists of probabilities is exploited in the determination of stability. In particular, starting with a given e, no perturbation dp^i_j can occur outside of the product simplex S.

B. Convergence Behavior of $\{p(n)\}_{n \geq 0}$

To show ϵ-optimality in the game Γ, it is shown that

$$\lim_{n \to \infty} \Pr\ \{p(n) = e^*\} > 1 - 0(a) \tag{A.8}$$

for any $p(0)$ in the interior of S. This is done in four steps.

i) We first establish that the collection of L_{R-I} schemes (3) used in Γ satisfy the definition of a distance-diminishing operator (5).

Let $\rho_k(n)$ be a particular value of $p(n)$ and the metric d be

$$d[\rho^i_1(n), \rho^i_2(n)] \triangleq \left(\sum_{j=1}^{r_i} [\rho^i_{1j}(n) - \rho^i_{2j}(n)]^2 \right)^{1/2}$$

where ρ^i_{kj} is the jth component of player A_i's probability vector at $\rho_k(n)$. Then

$$d[\rho_1(n), \rho_2(n)] = \left(\sum_{i=1}^{N} d^2[\rho^i_1(n), \rho^i_2(n)] \right)^{1/2}.$$

If at stage n the play $\boldsymbol{\alpha}$ is used, then from (3) it follows that

$$d[\rho_1(n+1), \rho_2(n+1)] = \begin{cases} (1-a)d[\rho_1(n), \rho_2(n)] & \text{with probability } d(\boldsymbol{\alpha}) \\ d[\rho_1(n), \rho_2(n)] & \text{with probability } 1 - d(\boldsymbol{\alpha}) \end{cases}.$$

Since $0 < a < 1$ and for any $\boldsymbol{\alpha}$, $0 < d(\boldsymbol{\alpha}) < 1$, the distance between $\rho_1(n)$ and $\rho_2(n)$ can remain unchanged only on a sample path of measure zero; otherwise, it must decrease.

The elements of V are stochastically and topologically closed and hence are ergodic kernels. Since the distance-diminishing property holds, $\{p(n)\}_{n \geq 0}$ is a compact Markov process whose only absorbing states are the elements of V [18]. Hence,

$$\lim_{n \to \infty} \Pr\ \{p(n) = e \in V\} = 1. \tag{A.9}$$

ii) Consider the deterministic differential equation

$$\dot{f}(t) = W[f(t)], \quad f(0) = p. \tag{A.10}$$

Let $f(n) \triangleq f(na)$ be a discrete approximation to $f(t)$ in (A.10). Then

$$\delta f(n) \triangleq f(n+1) - f(n) = aW[f(n)] + 0(a^2). \tag{A.11}$$

It can be shown using arguments given in [15] that

$$E[p(n) - f(n)] = k_1 a \tag{A.12}$$

and

$$E[\{p(n) - f(n)\}^2] = k_2 a, \text{ for any } p(0) \text{ and all } n = 0, 1, 2, \cdots.$$

Thus, the mean learning curve differs from the deterministic trajectory by only a small amount if the reward parameter a is small (see Fig. 1).

iii) Define $\bar{f}$ as a stationary point of (A.11). From the stability analysis in Part A of this Appendix, it follows that $f^* = e^*$ is the unique stable stationary point of (A.10), so that $W[e^*] = 0$. From (A.11), $a(W[\bar{f}] - W[e^*]) = 0(a^2)$ and hence

$$\bar{f} - e^* = 0(a). \tag{A.13}$$

From (A.12) and (A.13), $\lim_{n \to \infty} E[p(n)] - e^* = 0(a)$. But this can only be satisfied if

$$\lim_{n \to \infty} \Pr\ \{p(n) = e\} = k_e a \tag{A.14}$$

for some constant k_e and all $e \neq e^*$. Let $k_{\max} = \max_e \{k_e\}$. From (A.9) and (A.14), $\lim_{n \to \infty} \Pr\ \{p(n) = e^*\} > 1 - k_{\max} a$, which shows (A.8).

iv) Let $\boldsymbol{\alpha}_e$ be the play corresponding to the vertex $e \neq e^*$ of S. From the definition of $M(n)$ in (7), it follows that

$$\lim_{n \to \infty} E[M(n)] = d(\alpha^*) \left[1 - \sum_{e \neq e^*} \lim_{n \to \infty} \Pr\ \{p(n) = e\} \right] + \sum_{e \neq e^*} d(\boldsymbol{\alpha}_e) \lim_{n \to \infty} \Pr\ \{p(n) = e\}$$

$$= d(\alpha^*) - \sum_{e \neq e^*} [d(\alpha^*) - d(\boldsymbol{\alpha}_e)] k_e a.$$

Let

$$a^* = \frac{\epsilon}{\sum_{e \neq e^*} [d(\alpha^*) - d(\boldsymbol{\alpha}_e)] k_e}.$$

Then for any $a < a^*$, $\lim_{n \to \infty} E[M(n)] > d(\alpha^*) - \epsilon$, proving Theorem 1.

References

[1] R. A. Howard, *Dynamic Programming and Markov Processes.* Cambridge, MA: M.I.T. Press, 1960.

[2] S. M. Ross, *Applied Probability Models with Optimization Applications.* San Francisco, CA: Holden-Day, 1970.

[3] C. Derman, *Finite State Markovian Decision Processes.* New York: Academic, 1970.

[4] P. Mandl, "Estimation and control in Markov chains," *Adv. Appl. Prob.*, vol. 6, pp. 40-60, 1974.

[5] V. Borkar and P. Varaiya, "Adaptive control of Markov chains, I: Finite parameter set," *IEEE Trans. Automat. Contr.*, vol. AC-24, pp. 953-958, 1979.

[6] P. R. Kumar and W. Lin, "Optimal adaptive controllers for unknown Markov chains," *IEEE Trans. Automat. Contr.*, vol. AC-27, pp. 765-774, 1982.

[7] K. Astrom, "Optimal control of Markov processes with incomplete state information," *J. Math. Anal. Appl.*, vol. 10, pp. 174-205, 1965.

[8] K. Hsu and S. I. Marcus, "Decentralized control of finite state Markov processes," *IEEE Trans. Automat. Contr.*, vol. AC-27, pp. 426-431, 1982.

[9] R. E. Bellman, "A Markovian decision process," *J. Math. Mech.*, vol. 6, pp. 679-684, 1957.

[10] D. J. White, "Dynamic programming, Markov chains, and the method of successive approximations," *J. Math. Anal. Appl.*, vol. 6, pp. 373-376, 1963.

[11] P. Varaiya, "Optimal and suboptimal stationary controls for Markov chains," *IEEE Trans. Automat. Contr.*, vol. AC-23, pp. 388-394, 1978.

[12] J.-P. Forestier and P. Varaiya, "Multilayer control of large Markov chains," *IEEE Trans. Automat. Contr.*, vol. AC-23, pp. 298-305, 1978.

[13] Y. M. El-Fattah, "Recursive algorithms for adaptive contrul of finite Markov chains," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-11, pp. 135-144, 1981.

[14] M. Sato, K. Abe, and H. Takeda, "Learning control of finite Markov chains with unknown transition probabilities," *IEEE Trans. Automat. Contr.*, vol. AC-27, pp. 502-505, 1982.

[15] S. Lakshmivarahan, *Learning Algorithms: Theory and Applications.* New York: Springer-Verlag, 1981.

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

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

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

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

[20] M. L. Tsetlin, *Automaton Theory and Modeling of Biological Systems.* New York: Academic, 1973.

[21] K. S. Narendra and M. A. L. Thathachar, "Learning automata—A survey," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-4, pp. 323-334, 1974.

[22] K. S. Narendra and S. Lakshmivarahan, "Learning automata—A critique," *J. Cybern. Inf. Sci.*, vol. 4, pp. 53-66, 1977.

[23] S. Lakshmivarahan and M. A. L. Thathachar, "Absolute expediency of *Q*- and *S*- model learning algorithms," *Trans. Syst., Man, Cybern.*, vol. SMC-6, pp. 222-226, 1976.

[24] R. M. Wheeler, Jr. and K. S. Narendra, "Learning models for decentralized decision making," *Automatica*, vol. 21, pp. 479-484, 1985.

[25] S. Lakshmivarahan and K. S. Narendra, "Learning algorithms for two-person zero-sum stochastic games with incomplete information," *Math. Oper. Res.*, vol. 6, pp. 379-386, 1981.

[26] S. Lakshmivarahan and K. S. Narendra, "Learning algorithms for two-person zero-sum stochastic games with incomplete information: A unified approach," *SIAM J. Contr. and Opt.*, vol. 20, pp. 541-552, 1982.

[27] J. F. Nash, "Equilibrium points in n-person games," *Proc. National Acad. Sci. USA*, vol. 36, pp. 48-49, 1950.

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

[29] S. L. Ginsburg, V. Y. Krylov, and M. L. Tsetlin, "One example of a game for many identical automata," *Avtomat. Telemekh.*, vol. 25, pp. 668-671, 1964.

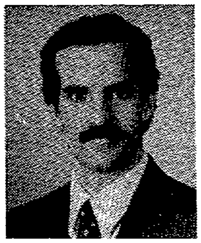
[30] V. I. Varshavskii, "Collective behavior and control problems," in *Machine Intelligence, 3,* D. Michie, Ed. Edinburgh: Edinburgh Univ., 1968.

[31] R. Viswanathan and K. S. Narendra, "Competitive and cooperative games of variable-structure stochastic automata," *J. Cybern.*, vol. 3, pp. 1-23, 1973.

[32] N. Baba and Y. Sawaragi, "On the learning behavior of stochastic automata under a nonstationary random environment," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-5, pp. 273-275, 1975.

[33] M. A. L. Thathachar and P. S. Sastry, "A new approach to the design of reinforcement schemes for learning automata," Dept. Elec. Eng., Ind. Inst. Sci., Bangalore, India, Tech. Rep. EE/60, 1983.

[34] 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, 1968.



Richard M. Wheeler, Jr. (M'84) received the B.S. degree in engineering and applied science from Yale University, New Haven, CT, in 1974, the M.S. degree in systems simulation and policy design from Dartmouth College, Hanover, NH, in 1980, and the Ph.D. degree in electrical engineering from Yale University in 1985.

He has worked on the application of mathematical modeling to energy conservation and small-scale hydropower development. At present, he is a Member of Technical Staff at Sandia Laboratories in Livermore, CA. His current research interests are in decentralized decision making under uncertainty, learning systems, and game theory.

Kumpati S. Narendra (S'55–M'66–SM'63–F'79) received the Ph.D. degree in applied physics from Harvard University, Cambridge, MA, in 1959.

He is currently the Director of the Center for Systems Science and the Chairman of the Department of Electrical Engineering, Yale University, New Haven, CT. He is the author (with J. H. Taylor) of the book *Frequency Domain Criteria for Absolute Stability* (New York: Academic, 1973), the editor (with R. V. Monopoli) of the book *Applications of Adaptive Control* (New York: Academic, 1980), and is currently editing a book entitled *Adaptive and Learning Systems: Theory and Applications* (New York: Plenum, 1986). His present interests are in adaptive control, learning theory, decentralized control of large scale systems, and control of flexible space structures.