Learning by Distributed Automata

Eric J. Friedman
Department of Industrial Engineering and Operations Research
University of California, Berkeley, CA 94720.

Scott Shenker
Xerox PARC
3333 Coyote Hill Road, Palo Alto, CA 94304

May 24, 1993

Abstract

We introduce a slightly modified version of a standard learning automaton and show that this 'responsive' learning automata eliminates dominated strategies when playing against an unknown environment. This new automaton has strong convergence properties that are easily analyzed, allowing us to compute explicit bounds for convergence rates.

Groups of such automata, interacting via a general game, are then studied. We show that synchronous groups of automata converge to the serially undominated set. We then show that, in contrast, asynchronous automata do not necessarily converge to the serially undominated set. However, these asynchronous automata do converge to the serially set-undominated set, which we define.

Preliminary Draft

1 Introduction

This paper combines the game theoretic approach to learning with the theory of decentralized control as exhibited by Learning Automata (LA). Our main results are (1) a modified version of a standard LA, called a responsive learning automata, which exhibits the properties of a 'rational' learner in that it eliminates dominated strategies, and (2) the application of several game theoretic properties of rational learning to describe the outcome of a game played by a group of such rational learners.

We consider learning in an environment where the players do not have any *a priori* information about the payoff function, and must rely on experimentation to determine which strategies to play. The key property of a rational learner is that when playing in an eventually fixed environment she will eventually learn to play the best strategy. Unfortunately, standard models of learning automata are not rational in this sense. For example, all 'absolutely expedient' LAs [10, 15] have the property that they may discard strategies. That is, after a certain time there is a finite probability that they will never play that strategy again. Thus if the environment changes and the discarded strategy becomes the best one, they will never notice or react to the change.

We remedy this deficiency by requiring that the probability of playing any specific strategy never goes below a fixed constant. Thus with probability one that strategy will be played infinitely often. This new learning automaton[1]turns out to be quite amenable to detailed analysis. For example, we can easily compute the expected convergence time explicitly. These 'responsive' learning automata (RLAs) are also quite well-behaved; they converge to the optimal strategy in a very strong sense.

Given these RLAs, we then apply a theory based on the work of Milgrom and Roberts [12, 13] to describe the behavior of groups of RLAs playing a game[2]. Our results show that if the RLAs play synchronously then they obey a generalized version of Milgrom and Roberts' 'consistency with adaptive learning' [13]. This implies that they will eventually be playing in the serially undominated set. However, when play is asynchronous, this is no longer true. In this case we show that they converge to the serial set-undominated set (to be defined) which is larger than the serial-undominated set.

This result is interesting as it shows that 'irrational' (in the game theoretic sense) results can occur when rational learners play asynchronously. It also has implications for decentralized control, an important use of learning automata [2, 11], and thus highlights the

[1]We have chosen a specific modification of a certain learning automata. However, we believe that most learning automata can be modified in a similar way, with similar results.

[2]Note that the description of multiple automata interacting as a game is very common in the learning automata literature [15, 8, 9].

difference between synchronous and asynchronous control. This is important in the design of decentralized systems. (See, for example [18].)

2 Set Limited Learning

In this section we describe the game theoretic model of interacting automata. The use of game theory as a paradigm for interacting automata is quite common. Perhaps the earliest example of this was in the work of Krylov and Tsetlin [7] where the two person zero-sum game is studied. Later, Lakshmivarahan and Narendra [8, 9] show that when two specific automata play a zero-sum game they will converge to the 'value' of the game, which is the standard game-theoretic outcome. Another important game that has been studied is the identical payoff game. This model is important for the decentralized optimization of a global quantity. (For a review of the identical payoff game see [15][pp.309–330].)

Results for more general games have been sparse, partly due to the lack of game-theoretic results. However, recent results in the study of learning for general games [12, 13, 3, 6] provide a framework for our study of interacting learning automata.

Consider a game [4] with n players. Assume that player a has m_a possible strategies $\Sigma_a = \{1, 2, \ldots, m_a\}$, and let $\Sigma = \Sigma_1 \times \cdots \Sigma_n$. Let s_a^t be player a's strategy at time t, s_{-a}^t be the strategies of all the other players, and $s^t = (s_a^t, s_{-a}^t)$. At time t, player a receives $\Gamma_a(s^t)$, where $\Gamma : \Sigma \mapsto \Re$ is the payoff function of the game.

We are interested in the eventual outcome of a game if players initially have no information about the payoff function, and are allowed to vary their strategies over time. Assume that one player is a learning automaton. Then the structure of her plays depends on what the other players are doing.

If the other players are playing strategies chosen randomly from a fixed probability distribution then it is as if she is playing against a random payoff function. This problem has been well studied [15]. Theorem 1 shows that our modified learning automata will almost always play the strategy with the highest expected payoff in this case.

Now, if the other players are not random, and they are learning over time, it is not clear what her asymptotic behavior should be. This asymptotic behavior, in fact will depend crucially on the assumptions we make about the other players. In much of the literature in economics [13, 3] it is assumed that players should either play a Nash equilibrium or, at the very least, a rationalizable equilibria [1, 16]. However, our automata, which do seem to be reasonable models of learning, do not necessarily converge to rationalizable equilibria. As we will show they may even converge to a Stackelberg equilibrium, even though we can show that they are reasonable, in a sense that we define below.

There are two important cases to consider. The first is that of non-predictive adversaries and synchronous games, where no player is allowed to see the other players' current strategies before playing her own. This will occur if all players are synchronized and are required to pick their strategies at the same time. In this case we show that players will hardly ever play dominated strategies. We say that strategy i dominates strategy j for player a if

$$\forall s_{-a}, \quad \Gamma_a(i, s_{-a}) > \Gamma_a(j, s_{-a})$$

That is, strategy i dominates strategy j for player a if, against any specific play by other players, the payoff for playing i is more than that for playing j. We can use the concept of domination to define a mapping on strategy sets. We define $U_a : 2^{\Sigma_{-a}} \mapsto \Sigma_a$ for any $S_{-a} \subseteq \Sigma_{-a}$, as the set of undominated strategies:

$$U_a(S_{-a}) = \{s_a \in \Sigma_a | \ \not\exists s'_a \in \Sigma_a \text{ s.t. } \forall s_{-a} \in S_{-a} \ \ \Gamma_a(s'_a, s_{-a}) > \Gamma_i(s_a, s_{-a})\}$$

Let $U = (U_1, \ldots, U_n)$. In the theory of games where it is usually assumed that players know the entire game matrix, it is a natural assumption that a rational player will not play a dominated strategy. This assumption does not apply in our context, because there is no knowledge of the payoff function and thus players cannot directly identify dominated strategies. However, we show that our automata do indeed eventually avoid playing dominated strategies and almost always play in $U_a(S_{-a})$.

The second case we consider is that of either predictive adversaries, or of non-predictive adversaries in an asynchronous game. Predictive adversaries can see the other players strategies before playing their own. An asynchronous game is where players vary their strategies at different times. In these cases of asynchrony or predictive adversaries, one cannot expect that players eventually play undominated strategies. In this situation we can only use a much weaker form of domination, which we call set-domination.

We say a strategy i for player a set-dominates another strategy j if all the possible payoffs associated with i exceed all those payoffs for j:

$$\min_{s_{-a} \in \Sigma_{-a}} \Gamma_a(i, s_{-a}) > \max_{s_{-a} \in \Sigma_{-a}} \Gamma_a(j, s_{-a})$$

Define $SU_a(S_{-a})$ to be the set of set-undominated strategies for player a, if all the other players are playing from the set $S_{-a} \subseteq \Sigma_{-a}$.

Set-domination is less restrictive than ordinary domination in that $U_a(S_{-a}) \subseteq SU_a(S_{-a})$. Set-domination is the appropriate concept when considering predictive adversaries. Even if we assuming that strategy i dominates strategy j, but another player always reacts to strategy i in a different way than they react to j, then it might turn out that it is in the player's best interest to play j. This would never be the case if i set-dominates j.

3 Responsive Learning Automata

In this section we define a slight variation on standard automata that have more resilient properties in a changing environment.

A typical learning automaton is given by Narendra and Thatcher [15]. Given an environment with m possible strategies and payoffs between 0 and 1, the automata consists of the vector of probabilities $p^t = (p_1^t, p_2^t, \ldots, p_m^t)$ where at time t the automata picks strategy i with probability p_i at random. Assuming that strategy i is picked with resulting payoff r_i^t, the probabilities are then updated in the following manner:

$$p_i^{t+1} = p_i^t + \alpha r_i^t (1 - p_i^t)$$

$$\forall j \neq i: \; p_j^{t+1} = p_j^t (1 - \alpha r^t)$$

Learning automata satisfying this rule are denoted LA_α. The automata LA_α does not satisfy our requirements for a 'good' learner. Against a static environment there is a high probability that it will eventually stop playing certain strategies; thus it would not notice if the payoffs for that strategy improved, thus rendering a previously bad strategy a good strategy to play.

We define a slight variation on LA_α that endows it with the properties of a good learner. Essentially, we require that no strategy ever has probability less than $\alpha/2$ of being played, thus it will be played infinitely often. The update rule for this automaton, denoted by RLA_α ($\alpha < 1/2$), is:

$$p_i^{t+1} = p_i^t + \alpha r_i^t \sum_{j \neq i} a_j^t p_j^t$$

$$\forall j \neq i \;\; p_j^{t+1} = p_j^t - \alpha r_i^t a_j^t p_j^t$$

where

$$a_j^t = \min[1, \frac{p_j^t - \alpha/2}{\alpha p_j^t r_i^t}]$$

Note that if all $p_j \geq \alpha$ then the update rule for RLA_α is the same as that for LA_α. We next show that this automaton is a good learner.

We define a random environment as one in which a player's opponents play randomly from a set of strategies. This model is general enough to encompass most random environments encompassed in the literature [15].

Definition 1 *Learning automaton a is playing against a random environment $S_{-a} \subseteq \Sigma_{-a}$ if at each t the strategy set $s_{-a}^t \in S_{-a}$ is chosen according to some (fixed) probability distribution.*

We will call a game Γ *normalized* if $\Gamma(s) \in [0,1]^n$ for all $s \in \Sigma$. Note that any game $\Gamma(s)$ can be easily transformed into a normalized game. Consider some player a. For each strategy $i \in \{1, 2, \ldots, m_a\}$, let r_i^t denote the random variable $\Gamma_a(i, s_{-a}^t)$. Thus, player a, when playing against a random environment, sees a fixed distribution of payoffs for each strategy i. In what follows, we therefore refer to expectation values for payoffs of a given strategy, using the notation $E[r_i^t]$ to denote the expected value of $\Gamma_a(i, s_{-a}^t)$:

$$E[r_i^t] = \sum_{\tilde{s}_{-a} \in S_{-a}} P[s_{-a}^t = \tilde{s}_{-a}]\Gamma_a(i, \tilde{s}_{-a}).$$

For random environments, these averages are independent of t so we can write $E[r_i]$.

Before stating our first convergence theorem, we need to define convergence.

Definition 2 *We say that a random variable x^t parametrized by α α-converges*[3] *to 0 if there exist positive constants α_0, β, b_1, b_2, b_3, and q such that, for any $0 < \alpha < \alpha_0$:*

- $\lim_{T\to\infty} \left(\frac{1}{T}\int_0^T dt P[x^t > \sqrt{\alpha}]\right) < \alpha$
- *If τ_f is the first time that $x^t \leq \beta\alpha$, then $E[\tau_f] \leq b_1/\alpha^q$.*
- *If τ_r is the first time that $x^t \geq \sqrt{\alpha}$, given that $x_\alpha^0 \leq \beta\alpha$, then $E[\tau_r] \geq b_2 e^{b_3/\sqrt{\alpha}}/\alpha$.*

Thus convergence is defined by a rapid collapse to optimality and a very long period at optimality before random variations cause inferior strategies to be played. Thus in any average the exponential part will dominate the polynomial.

Theorem 1 *Consider a normalized game Γ, some player a, and some random environment with support $S_{-a} \subseteq \Sigma_{-a}$. If $E[r_1] < E[r_2] \leq E[r_i]$ for all $i > 2$ then*

$$p_D^t \equiv \sum_{i=2}^n p_i^t$$

α-converges to 0 for any initial condition p^0.

Now we extend the concept of a random environment to allow for 'mistakes'.

[3]We note that this definition is not as sharp as possible for our model learning automaton. This is because we wish to emphasize that our results for mutiple automata are not overly dependent on our specific model of learning automaton. Thus our results for multiple automata should apply to any collection of automata that satisfy our basic convergence properties.

Definition 3 *We say that learning automaton a is playing against a 'δ-approximate' random environment S_{-a} if at each t s^t_{-a} is chosen from a random environment with probability greater than $1-\delta$, and randomly with probability less than δ a mistake may occur, that is $s^t_{-a} \notin S_{-a}$.*

Theorem 2 *Consider some normalized game Γ, some player a, some random environment with support $S_{-a} \subseteq \Sigma_{-a}$. If $E[r_1] < E[r_2] \leq E[r_i]$ for all $i > 2$ in this random environment then there exists a δ_0 such that for all $\delta \leq \delta_0$, if player a plays against a δ-approximate random environment, then*

$$p^t_D \equiv \sum_{i=2}^{n} p^t_i$$

α-converges to 0 for any initial condition p^0.

Consider now a non-predictive environment where the an adversary can pick any opposing strategy. but However he must do so withour prior knowlege of the automatons play except for the probabilities of play.

Definition 4 *Learning automaton a is playing against a non-predictive environment S_{-a} if at each t $s^t_{-a} \in S_{-a}$ where $S_{-a} \subseteq \Sigma_{-a}$, where s^t_{-a} is chosen without the knowledge of s^t_a.*

As for a random environment, we can extend the concept of a non-predictive environment to a δ-accurate non-predictive environment.

Definition 5 *We say that learning automaton a is playing against a 'δ-approximate' non-predictive environment S_{-a} if, at each t, s^t_{-a} is chosen from a non-predictive environment with probability greater than $1-\delta$, and randomly with probability less than δ a mistake may occur, that is $s^t_{-a} \notin S_{-a}$.*

This leads to the following theorem.

Theorem 3 *Consider some normalized game Γ, some player a and some non-predictive environment with support $S_{-a} \subseteq \Sigma_{-a}$. Then, there exists a δ_0 such that for all $\delta \leq \delta_0$, if player a plays against a δ-approximate non-predictive environment with support S_{-a}, then*

$$p^t_D \equiv \sum_{i \notin U(S_{-a})} p^t_i$$

α-converges to 0 for any initial condition p^0.

Against a possibly predictive adversarial environment we can only prove a weaker statement.

Definition 6 *Learning automaton i is playing against a environment S_{-a} if at each t $s_{-a}^t \in S_{-a}$ with $S_{-a} \subseteq \Sigma_{-a}$, and where s_{-a}^t may be chosen with the knowledge of s_a^t.*

Theorem 4 *Consider some normalized game Γ, some player a and some environment with support $S_{-a} \subseteq \Sigma_{-a}$. Then, there exists a δ_0 such that for all $\delta \leq \delta_0$, if player a plays against a δ-approximate environment with support S_{-a}, then*

$$p_D^t \equiv \sum_{i \notin SU(S_{-a})} p_i^t$$

α-converges to 0 for any initial condition p^0.

4 Multiple Learning Automata

Consider a game that is being played continuously in time. Each player can at any time change her strategy or evaluate the success (payoff) of her current strategy. For example, consider several users sharing a network link. At each instant each user has a certain link utilization. At any time a user can change her utilization. Then she may compute the success of the current utilization level as some average over a certain amount of time. This will be our model of a learning automaton playing a continuous time game.

First we will consider the case where all the automata update their strategies at the same time.

4.1 Synchronous Automata

In this case we imagine that time is discrete, and all the automata update their strategy with each unit of time. However, they may all have different α's, subject to the mild restriction[4] that $\alpha_{max}^p < \alpha_{min}$, where α_{max} is the largest α and α_{min} the smallest.

Theorem 5 *For any group of n synchronous learning automata RLA_α, $n > 1$ and $p \geq 1$ there exists some α_0 such that if all the learning automata in the group have $\alpha \leq \alpha_0$ and $\alpha_{max}^p < \alpha_{min}$, then for any automaton*

$$p_D^t \equiv \sum_{i \notin U_a^\infty(\Sigma)} p_i^t$$

[4]We believe that this restriction could be removed, with a more detailed analysis.

α-converges to 0, where α-convergence is defined as all α_a's converge to zero while satisfying $\alpha_{max}^p < \alpha_{min}$.

The set $U^\infty(\Sigma)$, the result of the iterated elimination of strictly dominated strategies, has been much studied in economic learning theory[13]. Many important learning models have been shown to converge there. In fact a very large class of games, those which are supermodular, or have strategic complementarities, this set is very simple. For example, both the 'General Equilibrium Model with gross substitutes' and the Bertrand oligopoly model with differentiated products have a singleton $U^\infty(\Sigma)$. Thus in these (and other) important economic models synchronous learning automata converge to the unique, and in some sense 'correct', equilibrium.

This result only requires that the learning automata satisfy our Lemmas 1 and 2 in the Appendix. Thus, these results should hold quite generally, for large classes of learning automata.

4.2 Asynchronous Automata

Now consider the case where time is no longer discrete and each automaton independently chooses when to change strategies. In this case it is not clear what the 'correct' method for determining the payoff of a particular strategy, so we allow for a wide variety of possibilities.

Let $RLA_\alpha^{T,G}$ be a learning automata which updates its strategy every T units of time. The payoff that it uses for its update is some weighted average of its payoffs in the previous time period; if player a has been playing strategy i for the past time period then the reward is

$$r_i^t = \frac{1}{T}\int_{t-T}^{t} \Gamma_a(s^{t'})dG(t'-t)$$

where $G(t)$ is a monotone function, $s_a^{t'} = i$ for all $t' \in [t-T, t]$, and we interpret the above expression as a Riemann-Stieltjes integral.

For example, if $G(t) = t/T$ then

$$r_i^t = \frac{1}{T}\int_{t-T}^{t} \Gamma_a(s^{t'})dt'$$

which is just the average.

Another useful choice for G is:

$$G(t) = 0 \quad t \le T - b$$

$$G(t) = 1 \quad t > T - b$$

with $0 \leq b \leq T$. In this case $r_i^t = \Gamma_a(s^{t-b})$. This can obviously be generalized to any pointwise average.

Any group of such learning automata with different T's we shall call asynchronous.

Theorem 6 *For any group of n asynchronous learning automata RLA_α, $n > 1$ and $p \geq 1$ there exists some α_0 such that if all the learning automata in the group have $\alpha \leq \alpha_0$ and $\alpha_{max}^p < \alpha_{min}$, where α_{max} is the largest α and α_{min} the smallest, then for any automaton*

$$p_D^t \equiv \sum_{i \notin SU_a^\infty(\Sigma)} p_i^t$$

α-converges to 0. (Where α-convergence is defined as all α_a's converge to zero while satisfying $\alpha_{max}^p < \alpha_{min}$).

In this case the result is not a strong as one would desire. For many important games $SU^\infty(\Sigma)$ is not a singleton, and then our theorem does not uniquely define the outcome. However, this is necessary, as the specific outcome is dependent on the timing and averaging of the different automata.

For example, one possible outcome is a Stackelberg equilibrium [4]. This is interesting from the game theoretic viewpoint, as this is not a possible outcome in standard models of economic learning theory [13]. This outcome occurs in a two automata game when the first automaton (A1) is updating much more often than the second (A2) . Then since A1 is always going to the best reply to A2's strategy, we see that A2 is Stackelberg leader, and they will converge to the Stackelberg equilibrium.

Theorem 7 *In the two player normalized game there exist $RLA_{\alpha_1}^{T_1,G_1}$ versus $RLA_{\alpha_2}^{T_2,G_2}$ such that player 1 converges to Stackelberg leader and player 2 to follower[5].*

This is interesting as in many games (i.e. all those that have a pure Nash equilibrium) players would prefer to be leader than follower. Thus A1 is doing worse by updating often than if he were updating very rarely. This seems counter-intuitive, as one would expect that rapid response would be a desirable attribute.

However, certain games are 'non-manipulable', in that $SU^\infty(\Sigma)$ is a singleton, and thus any set of asynchronous automata (with small enough α's) will converge to a unique strategy.

We now define class of games that have this property. Following [14] we define the class of *generalized serial* games to be those that have the following five properties:

[5]This can be easily generalized to the multi-player Stackelberg equilibria.

- Ordered strategy domains: $\Sigma_a \subseteq \Re$
- Cross-Monotonicity: $\Gamma_a(s) \geq \Gamma_i(\tilde{s}_b, s_{-b})$ for any $\tilde{s}_b \geq s_b$
- Seriality: $\Gamma_a(s_b, s_{-b}) = \Gamma_a(\tilde{s}_b, s_{-b})$ for any $s_b, \tilde{s}_b \geq s_a$
- Unique best reply: for each s_{-a} there exists an element $BR_a(s_{-a})$ such that $x_a \neq BR_a(s_{-a}) \Rightarrow \Gamma_a(BR_a(s_{-a}), s_{-a}) > \Gamma_a(x_a, s_{-a})$
- Seriality of best reply: $BR_a(s_{-a}) = BR_a(\tilde{s}_b, s_{-b})$ for any $\tilde{s}_b \geq BR_a(s_{-a})$

Theorem 8 *Generalized serial games have a singleton $SU^{\infty}(\Sigma)$.*

Generalized serial games arise from the division of output in a production economy or the sharing of an externality [14]. A relevant example of the latter is the sharing of communication link operating under the Fair-Share service discipline [17, 19]. If all the users of the link were using RLAs to compute their optimal transmission rate, then even under asynchronous updating and variations of α, the users will converge to the serial-set-undominated set, which in this case is a Pareto optimal Nash equilibrium [14]. Note that if the same communication link was using the standard first-in-first-out (FIFO) discipline then convergence might not occur. However, in this case if the RLAs were synchronous then convergence to Nash would occur. Thus there are important differences between synchronous and asynchronous automata that must be taken into account when they are used for decentralized control.

5 Conclusions

We have presented a modification of a standard learning automata which endows it with the properties of a rational learner. This new automaton has good convergence properties, many of which can be explicitly calculated.

Groups of these automata that interact via a game have a well defined behavior. Synchronous automata converge to the serial-undominated set while asynchronous will converge to the serial-set-undominated set.

While we have restricted our presentation to a single model of a learning automaton, we note that these results should extend to many other models, as the proof techniques that we apply are quite general. In fact we believe that these methods could be used to simplify many of the results in standard learning automata theory. Additionally, most of these results extend naturally to the case of stochastic games. Thus even in games with a random component they apply.

Finally, we comment that these results have implications for the game theoretic study of learning, since supposedly 'rational' behavior can lead to a larger class of behavior than often assumed. We hope to extend our work to the case of continuous games, and expect that similar behavior will appear.

6 Acknowledgements

We would like to thank Sendhil Mullainathan for numerous discussions on the nature of rational learning. We also thank David Aldous and Sheldon Ross for useful conversations. EJF would like to acknowledge the support of the Xerox Palo Alto Research Center where much of this work was done, and National Science Foundation Grant NSF-IRI-8902813.

A Appendix

We start by proving two useful lemmas.

A.1 Lemma 1

We consider some automata with a set of payoffs r_i^t that can either be random, or deterministic.

Lemma 1 *Consider some set of strategies A and define $p_A^t = \sum_{i \in A} p_i^t$. Assume there exists some $\beta > 1$ such that $E[r_i^t] < (1 - \frac{n}{\beta})E[r_j^t]$ for all t, for all $i \in A$, and for all $j \notin A$. Then, p_A^t α-converges to 0.*

We prove this lemma with the following sequence of claims. Choose some $k > 2\beta$. Define $r_A^t = \max_{i \in A} E[r_i^t]$ and $r_{-A}^t = \min_{i \notin A} E[r_i^t]$.

Claim 1 *There exists some constant c_1 such that if $p_A^t > \beta\alpha$ then*

$$E[p_A^{t+1} | p_A^t] \leq p_A^t - c_1\alpha^2.$$

Proof: Computing directly from the updating equations,

$$E[p_A^{t+1} | p_A^t] = p_A^t + \alpha \sum_{j \notin A} \sum_{i \in A} p_j p_i (E[min[r_i^t, \frac{p_j^t - \alpha/2}{\alpha p_j^t}]] - E[min[r_j^t, \frac{p_i^t - \alpha/2}{\alpha p_i^t}]])$$

Clearly

$$\sum_{i\in A} p_i E[min[r_i^t, \frac{p_j^t - \alpha/2}{\alpha p_j^t}]] \le p_A^t r_A^t$$

and

$$\sum_{i\in A} p_i E[min[r_j^t, \frac{p_i^t - \alpha/2}{\alpha p_i^t}]] \ge (p_A^t - n\alpha) r_{-A}^t$$

Combining, we have

$$E[p_A^{t+1}|p_A^t] - p_A^t \le \alpha p_A^t (1-p_A^t)(r_A^t - (1-\frac{n\alpha}{p_A^t}) r_{-A}^t)$$

Since $\frac{p_A^t(1-p_A^t)}{\alpha} > 1/4$ and $(r_A^t - (1-\frac{n\alpha}{p_A^t})r_{-A}^t) < 0$ we have proven the claim.$\diamond$

Claim 2 *Let τ_f be the first time that $p_A^t \le \beta\alpha$. Then*

$$E[\tau_f] < \frac{1}{c_1\alpha^2}$$

Proof: This proof follows that in [5]. Define

$$q^t = p_A^{\min(t,\tau_f)} + c_1\alpha^2 \min(t,\tau_f)$$

This is a submartingale

$$E[q^{t+1}|q^t] \le q^t$$

so

$$E[q^t] \le p_A^0$$

and thus

$$c_1\alpha^2 E[\min(t,\tau_f)] \le p_A^0$$

Taking the limit at $t \to \infty$ we find that

$$\lim_{t\to\infty} c_1\alpha^2 E[\min(t,\tau_f)] \le p_A^0$$

and

$$\lim_{t\to\infty} c_1\alpha^2 E[\min(t,\tau_f)] = E[\lim_{t\to\infty} c_1\alpha^2 \min(t,\tau_f)] = c_1\alpha^2 E[\tau_f]$$

by the monotone convergence theorem. Therefore,

$$E[\tau_f] \le \frac{p_A^0}{c_1\alpha^2} < \frac{1}{c_1\alpha^2}$$

$\diamond$

Now let $\hat{p}^t_A$ denote the process which is the stopped version of p^t_A where stopping occurs as soon as $p^t_A < \beta\alpha$ or $p^t_A > k\alpha$.

Claim 3 *There exists some $c_3 > 0$ such that $e^{c_3 \hat{p}^t_A/\alpha}$ is a submartingale.*

Proof: Let

$$z^t = e^{c\hat{p}^t_A/\alpha}$$

for some constant $c > 0$. The submartingale condition

$$E[z^{t+1}|z^t] \leq z^t$$

requires that

$$E[e^{c(\hat{p}^{t+1}_A - \hat{p}^t_A)/\alpha}] = \frac{E[z^{t+1}|z^t]}{z^t} \leq 1$$

Clearly, when $\hat{p}^t_A < \beta\alpha$ or $\hat{p}^t_A > k\alpha$ then

$$E[z^{t+1}|z^t] = z^t$$

For $\beta\alpha \leq \hat{p}^t_A \leq k\alpha$, with probability $\hat{p}^t_A$, for some $i \in A$

$$\hat{p}^{t+1}_A - \hat{p}^t_A = \sum_{j \notin A} p^t_j \min[r^t_i, \frac{p^t_j - \alpha/2}{\alpha p^t_j}] \leq \alpha(1 - \hat{p}^t_A) r^t_A$$

and with probability $1 - \hat{p}^t_A$, for some $j \notin A$,

$$\hat{p}^{t+1}_A - \hat{p}^t_A = \alpha \sum_{i \in A} p_i \min[r^t_j, \frac{p^t_i - \alpha/2}{\alpha p^t_i}] \leq -\alpha(\hat{p}^t_A - n\alpha) r^t_{-A}$$

Thus,

$$f(c) \equiv E[e^{c(p^{t+1}_A - p^t_A)/\alpha}] \leq p^t_A e^{c(1-p^t_A)r^t_A} + (1 - p^t_A) e^{-cp^t_A(\hat{p}^t_A - n\alpha) r^t_{-A}}$$

Note that $f(0) = 0$ and $f'(0) < 0$, so there exists some constant c_3 such that $f(c_3) < 0$. For this constant, z^t is a submartingale.$\diamond$

Claim 4 *If $\hat{p}^0_A \leq 2\beta\alpha$ then there exists some constant c_4 such that*

$$P[\lim_{t \to \infty} \hat{p}^t_A > k\alpha] < c_4 e^{-c_3 k}$$

Proof: Let P_k^t be the probability that $\hat{p}_A^t > k\alpha$, P_f^t be the probability that $\hat{p}_A^t < \beta\alpha$, $P_k = \lim_{t\to\infty} P_k^t$, and $P_f = \lim_{t\to\infty} P_f^t$. Let

$$z^t = e^{c_3 \hat{p}_A^t/\alpha}$$

For all t we have

$$E[z^t] = E[z^t|z^t < e^{\beta c_3}]P_f^t + E[z^t|z^t > e^{kc_3}]P_k^t + E[z^t|e^{\beta c_3} \leq z^t \leq e^{kc_3}](1 - P_f^t - P_k^t)$$

Thus,

$$\begin{aligned} P_k(E[z^t|z^t > e^{kc_3}] - E[z^t|z^t < e^{\beta c_3}]) = E[z^t] - E[z^t|z^t < e^{\beta c_3}] + \\ (P_f - P_f^t)(E[z^t|z^t < e^{\beta c_3}] - E[z^t|e^{\beta c_3} \leq z^t \leq e^{kc_3}]) + \\ (P_k - P_k^t)(E[z^t|z^t > e^{kc_3}] - E[z^t|e^{\beta c_3} \leq z^t \leq e^{kc_3}]) \end{aligned}$$

Thus, upon taking the limit $t \to \infty$,

$$P_k = \frac{E[z^t] - E[z^t|z^t < e^{\beta c_3}]}{E[z^t|z^t > e^{kc_3}] - E[z^t|z^t < e^{\beta c_3}]}$$

Since z^t is a submartingale we know that $1 \leq E[z^t] \leq z^0 < e^{2\beta c_3}$. Also, $1 \leq E[z^t|z^t < e^{\beta c_3}] \leq e^{\beta c_3}$, $e^{kc_3} \leq E[z^t|z^t > e^{kc_3}] \leq e^{(k+1)c_3}$, and $e^{\beta c_3} \leq E[z^t|e^{\beta c_3} \leq z^t \leq e^{kc_3}] \leq e^{kc_3}$. Thus,

$$P_k \leq e^{-kc_3} \frac{e^{2\beta c_3} - 1}{1 - e^{-(k-1)c_3}}$$

$\diamond$

Claim 5 *Assume that $p_A^0 \leq 2\beta\alpha$. Let τ_k be the first time for which $p_A^\tau \geq k\alpha$. Then*

$$E[\tau_k] > \frac{e^{c_3 k}}{2c_4\alpha}$$

Proof: Since

$$E[p_A^{t+1} - p_A^t | p_A^t \leq 2\beta\alpha] \leq 2\beta\alpha^2$$

the expected time to go from $p_A^t \leq \beta\alpha$ to $p_A^{t'} \geq 2\beta\alpha$ is at least $1/2\alpha$. Thus the expected time until $p_A^\tau \geq k\alpha$ is

$$E[\tau_k] \geq E[\text{number of times to } \alpha \text{ before } k\alpha]/2\alpha = \frac{1}{2\alpha P_k} > \frac{e^k}{2c_4\alpha}$$

$\diamond$

Claim 6 *Assume that $p_A^0 \leq 2\beta\alpha$. Let τ_r be the first time that $p_A^t \geq \sqrt{\alpha}$. Then,*

$$E[\tau_r] \geq \frac{e^{c_3/\sqrt{\alpha}}}{2c_4\alpha}$$

Proof: This follows immediately from choosing $k = 1/\sqrt{\alpha}$ in the preceding claim.$\diamond$

Claim 7 *There exists some α_0 such that for all $\alpha < \alpha_0$*

$$\lim_{T\to\infty} \frac{1}{T}\int_0^T dt P[p_A^t > \sqrt{\alpha}] < \alpha$$

Consider the process with three states: A:$p_A^t < \beta\alpha$, B:$p_A^t \geq \beta\alpha$ and $p_A^{t'} < \beta\alpha$ has occured more recently than $p_A^{t'} > \sqrt{\alpha}$, and C:$p_A^t \geq \beta\alpha$ and $p_A^{t'} > \sqrt{\alpha}$ has occured more recently than $p_A^{t'} < \beta\alpha$. The system goes from state A to state B to state C and then back to state A. The expected time to make a transition from A to B is bounded below by 1. The expected time to make a transition from B to C is bounded below by $E[\tau_r]$. The expected time to make a transition from C to A is bounded above by $E[\tau_f]$. Thus, the fraction of the time spent in state C (which is an upper bound on the averaged probability that $p_A^t > \sqrt{\alpha}$) is bounded above by

$$\frac{E[\tau_f]}{E[\tau_f] + E[\tau_r]} < \frac{E[\tau_f]}{E[\tau_r]} < e^{-c_3/\sqrt{\alpha}}\frac{2c_4}{c_1\alpha}$$

For small enough α_0, $e^{-c_3/\sqrt{\alpha}}\frac{2c_4}{c_1\alpha} < \alpha$ for all $\alpha < \alpha_0$. $\diamond$

Setting α_0 and β as above, and setting $b_1 = 1/c_1$, $b_2 = 1/2c_4$, and $b_3 = c_3$, we see that we have established the α-convergence.$\square$

A.2 Lemma 2

Lemma 2 *Consider an automata RLA_α playing against an environment with a set of payoffs r_i^t, and the same automata playing against a different environment with a set of payoffs $\hat{r}_i^t$; let p_i^t and $\hat{p}_i^t$ denote the probabilities in the two cases. Let A be any set of strategies. Then, if $\hat{r}_i^t$ is stochastically greater than or equal to r_i^t for all t and all $i \in A$ and if $\hat{r}_i^t$ is stochastically less than or equal to r_i^t for all t and for all $i \notin A$, then $\sum_{i\in A}\hat{p}_i^t$ stochastically dominates $\sum_{i\in A} p_i^t$ for all t.*

Proof: Define $p_A^t = \sum_{i\in A} p_i^t$. Notice that the update rules for p_A^t are, when strategy i is chosen at step t,

$$i \in A: \qquad p_A^{t+1} = p_A^t + \alpha r_i^t \sum_{j\notin A} a_j^t p_j^t$$

$$i \notin A: \qquad p_A^{t+1} = p_A^t - \alpha r_i^t \sum_{j \in A} a_j^t p_j^t$$

where

$$a_j^t = \min[1, \frac{p_j^t - \alpha/2}{\alpha p_j^t r_i^t}]$$

Thus, p_A^{t+1} is monotonically increasing in p_A^t, monotonically increasing in r_i^t with $i \in A$, and monotonically decreasing in r_i^t with $i \notin A$. Now consider the set of sample paths where a uniform random number in $[0, 1]$ is chosen at each iteration. It is easy to see that over any sample path $\hat{p}_A \geq p_A$. $\Box$

A.3 Proof of Theorem 1

Here, the payoffs r_i^t are random variables with distributions that are independent of t. Setting $A = \{1\}$ we can then apply Lemma 1 directly.$\Box$

A.4 Proof of Theorem 2

We start with some normalized game Γ, some player a and some non-predictive environment with support $S_{-a} \subseteq \Sigma_{-a}$. Consider the game $\hat{\Gamma}(s)$ defined as: $\hat{\Gamma}_a(s_a, s_{-a}) = \Gamma_a(s_a, s_{-a})$ for all $s_{-a} \in S_{-a}$, $\hat{\Gamma}_a(1, s_{-a}) = \min_{\tilde{s}_{-a} \in \Sigma_{-a}} \Gamma_a(1, \tilde{s}_{-a})$ for all $s_{-a} \notin S_{-a}$, and $\hat{\Gamma}_a(i, s_{-a}) = \max_{\tilde{s}_{-a} \in \Sigma_{-a}} \Gamma_a(1, \tilde{s}_{-a})$ for all $s_{-a} \notin S_{-a}$ and $i > 1$. Notice that in this new game, since when $s_{-a} \notin S_{-a}$ the payoffs are independent on the exact choice of s_{-a}, we need not concern ourselves with the nature of play outside of the random environment; we need only consider the rate at which 'mistakes' are made. With perhaps relabelling some strategies i with $i > 1$, we can find some δ_0 such that for all $\delta \leq \delta_0$ the δ-approximate environment has the property that $E[r_1] < E[r_2] \leq E[r_i]$ for all $i > 2$. Furthermore, whatever the play outside of S_{-a}, the payoffs in this new game for playing strategy 1 are stochastically smaller than those in the original game, and the payoffs in this new game for playing strategies other than 1 are stochastically greater than those in the original game. Therefore, theorem 2 follows from Theorem 1 and Lemma 1.$\Box$

A.5 Proof of Theorem 4

Here, the plays s_{-a}^t are not independent of the plays s_a^t. Thus, we will write $s_{-a}^t(s_a^t)$ to denote this dependence. Let D be the set of set-dominated strategies, U the set of set-dominating strategies, and M the remaining strategies. Consider the following payoffs. Define

$$r_D^t = \max_{s_a \in D} \max_{s_{-a} \in S_{-a}} \Gamma_a(s_a, s_{-a}) \; if \; s_{-a}^t(s_a^t) \in S_{-a}$$

$$r_D^t = \max_{s_a \in D} \max_{s_{-a} \in \Sigma_{-a}} \Gamma_a(s_a, s_{-a}) \; if \; s_{-a}^t(s_a^t) \notin S_{-a}$$

$$r_U^t = \min_{s_a \in U} \min_{s_{-a} \in S_{-a}} \Gamma_a(s_a, s_{-a}) \; if \; s_{-a}^t(s_a^t) \in S_{-a}$$

$$r_U^t = \min_{s_a \in U} \min_{s_{-a} \in \Sigma_{-a}} \Gamma_a(s_a, s_{-a}) \; if \; s_{-a}^t(s_a^t) \notin S_{-a}$$

$$r_M^t = \min_{s_a \notin U} \min_{s_{-a} \in S_{-a}} \Gamma_a(s_a, s_{-a}) \; if \; s_{-a}^t(s_a^t) \in S_{-a}$$

$$r_M^t = \min_{s_a \notin U} \min_{s_{-a} \in \Sigma_{-a}} \Gamma_a(s_a, s_{-a}) \; if \; s_{-a}^t(s_a^t) \notin S_{-a}$$

Note that whenever $s_{-a}^t(s_a^t) \in S_{-a}$ we have $r_U^t > r_D^t \geq r_M^t$. Furthermore, $r_i^t \geq r_U^t$ for all $i \in U$, $r_i^t \leq r_D^t$ for all $i \in D$, and $r_i^t \geq r_M^t$ for all $i \in M$. Consider the game where, at each time t if strategy i is played we assign the payoff r_U^t if $i \in U$, r_M^t if $i \in M$, and r_D^t if $i \in D$. In this new game, $\hat{p}_D^t$ stochastically dominates p_D^t by Lemma 2. Furthermore, if we look at the set $A \equiv D \cup M$, then we can apply theorem 2 to p_A^t to see that is α-converges to 0.$\square$

A.6 Proof of Theorem 3

Here, the plays s_{-a}^t are independent of the plays s_a^t. The proof for Theorem 3 is a slightly more complicated variation of Theorem 4. In this case we must eliminate strategies one at a time. Let strategy i dominate strategy j, and define M to be the set of remaining strategies.

$$r_j^t = \Gamma_a(j, s_{-a}^t) \; if \; s_{-a}^t \in S_{-a}$$

$$r_j^t = \max_{s_{-a} \in \Sigma_{-a}} \Gamma_a(j, s_{-a}) \; if \; s_{-a}^t \notin S_{-a}$$

$$r_i^t = \Gamma_a(i, s_{-a}^t) \; if \; s_{-a}^t \in S_{-a}$$

$$r_i^t = \min_{s_{-a} \in \Sigma_{-a}} \Gamma_a(i, s_{-a}) \; if \; s_{-a}^t \notin S_{-a}$$

$$r_M^t = \min_{s_a \neq i} \Gamma_a(s_a, s_{-a}^t) \; if \; s_{-a}^t \in S_{-a}$$

$$r_M^t = \min_{s_a \neq i} \min_{s_{-a} \in \Sigma_{-a}} \Gamma_a(s_a, s_{-a}) \; if \; s_{-a}^t(s_a^t) \notin S_{-a}$$

Consider the game where, at each time t if strategy i is played we assign the payoff r_i^t, if strategy j is played we assign the payoff r_j^t, and if any other strategy is played we assign r_M^t. In this new game, $\hat{p}_j^t$ stochastically dominates p_j^t by Lemma 2. Furthermore, if we look at the set $A \equiv i \cup M$, then we can apply theorem 2 to p_A^t to see that is α-converges to 0.$\square$

A.7 Proof of Theorems 5 and 6

Theorem 5 follows from the repeated application of theorem 3. For example initially theorem 3 requires that all players collapse down to the undominated set $S^1 = U(\Sigma)$. Then as all players are in S^1, theorem 3 now implies that they will collapse down to $S^2 = U^2(\Sigma)$. This process continues until they are all in $S^\infty = U^\infty(\Sigma)$. The same proof applies to theorem 5 where we replace U by SU, theorem 3 by theorem 4, and take into account the different time intervals. We will present the proof for theorem 6, and comment that theorem 5 is proved in the same manner.

First note that there exists an N such that $SU^N(\Sigma) = SU^\infty(\Sigma)$ as Γ is a finite game. Choose γ such that

$$(1-\gamma)^{Nn} \geq 1/2$$

We will show that the probability of a collapse in time

$$T_h N c_1 / \gamma \alpha_0^{3p}$$

is greater than $1/2$ where T_h is the largest update time among the automata.

Claim 8

$$Pr[\tau_f \leq \frac{T_h N c_1}{\gamma \alpha_0^{3p}}] \geq 1/2$$

Proof: Consider $S^m = SU^m(\Sigma)$ for $0 \leq m \leq N$. Let δ_a^m be the δ_0 required in theorem 3 for a δ-approximate environment with support set S_{-a}^m. Also choose $\hat{\alpha}_a^m$ to be the α_0 required by automaton a against the same environment. Let $\delta_0 = \min_{a,m} \delta_a^m$ and $\hat{\alpha}_0 = \min_{a,m}[\hat{\alpha}_a^m, \delta_0^2]$.

Now choose $\alpha_0 \leq \hat{\alpha}_0$ such that for all $\alpha \leq \alpha_0$ satisfying the restriction $\alpha_{max}^p \leq \alpha_{min}$ the following holds,

$$T_l c_2 e^{c_3/\sqrt{\alpha_{max}}} / \alpha_{max} > \frac{T_h N c_1}{\gamma \alpha_{min}^2}$$

where c_1 is the largest and c_2, c_3 the smallest of the c's that occur in theorem 3 against the different environments mentioned above and T_l is the smallest update time among the automata. By the restriction on the α's we see that this reduces to

$$T_l c_2 e^{c_3/\sqrt{\alpha_0}} / \alpha_0 > \frac{T_h N c_1}{\gamma \alpha_0^{2p}}$$

which guarantees that such an α_0 exists.

Now if all automata have $\alpha \in [\alpha_0^p, \alpha_0]$ the above construction guarantees that dominated strategies will never get large during the N repeated actions of the domination operator.

Thus with probability less than $\hat{\delta}_0$ all automata will be playing with $p_i^t \leq \sqrt{\alpha_0}$ for $i \notin SU(S_{-j}^m)$ when

$$mb_1/(\gamma\alpha_0^{2p}) \leq t \leq b_2 e^{b_2/\sqrt{\alpha}}/\alpha_0$$

This is the probability that an iteration of the domination operator will occur properly.

Thus by our definition of γ the collapse to $SU^\infty(\Sigma)$ with probability greater 1/2 will occur in the specified time. $\diamond$

Claim 9

$$E[\tau_f] \leq \frac{2T_h N c_1}{\gamma\alpha_0^{3p}}$$

Proof: τ_f is bounded above by a random variable with a geometric distribution, and the expected number of periods of length

$$\frac{T_h N c_1}{\gamma\alpha_0^{3p}}$$

is 2. $\diamond$.

Thus we have shown that the collapse will occur. Then the probabilities will remain small for an exponential (in α) amount of time by the α-convergence of the individual automata. $\square$

A.8 Proof of Theorem 7

Let α_1, δ_1 be such that theorem 1 applies against the game $\Gamma_1(s_1, BR(s_1))$ where $BR(s_1) = \text{argmax}_{s_2}\Gamma(s_1, s_2)$, where this environment is δ approximate, i.e. occasionally s_2 is not $BR(s_1)$. Let α_2 similarly defined for any $s_1 \in \Sigma_1$ for the game $\Gamma_2(s_1, s_2)$ played by automaton 2. Now set $T_2 = 1$ and $T_1 = 2K/\alpha_2^2$, where K is chosen sufficiently large so that after time T_1 the probability that player 2 is not playing $BR(s_1)$ is less than δ_0. Let $G_1 = G_{T_1}$ and $G_2 = t/T_2$, thus player 1 samples at the end of her interval, and player 2 averages over his entire interval.

Thus for any strategy that automaton 1 plays, player 2 will be at best reply with probability greater that $1 - \delta_0$ when player 1 samples for the value of his strategy. Thus player 1 observes the above defined game with at most δ_0 mistakes. If δ_0 is small enough then player 1 will converge to the best strategy, thus he will become Stackelberg leader.

Player 2 will then face a δ approximate environment with s_1 fixed at Stackelberg leader, thus she will converge to Stackelberg follower. $\square$

A.9 Proof of Theorem 8

Since the SU operator is monotonic, the iteration process must converge to a nontrivial fixed point. Let this fixed point of SU be denoted by $I = (I_1, I_2, \ldots, I_n)$ with $\bot_a$ denoting the minimal element of I_a and $\top_a$ denoting the maximal element of I_a, and $\bot$ and $\top$ denoting the vectors of these extremal elements. Let $MAX_a(x_a) = max_{s_{-a} \in I_{-a}} \Gamma_a(x_a, s_{-a})$, and $MIN_a(x_a) = min_{s_{-a} \in I_{-a}} \Gamma_a(x_a, s_{-a})$. For any $s \in I$ and for any $x_a \in I_a$, $\Gamma_a(x_a, \top_{-a}) \leq \Gamma_a(x_a, s_{-a}) \leq \Gamma_a(x_a, \bot_{-a})$ so $MAX_a(x_a) = \Gamma_a(x_a, \bot_{-a})$ and $MIN_a(x) = \Gamma_a(x_a, \top_{-a})$. Assume that I is not a singleton, so the set $\{a | \bot_a < \top_a\}$ is nonempty. We can define a as the element in this set with the smallest $\bot_a$: $\bot_b < \top_b \Rightarrow \bot_b \geq \bot_a$. In particular, $\Gamma_a(\bot_a, \top_{-a}) = \Gamma_a(\bot)$, so $MIN_a(\bot_a) = MAX_a(\bot_a)$. If there exists some $x_a \in I_a - \bot_a$ such that $\Gamma_a(x_a, \bot_{-a}) < \Gamma_a(\bot)$, then $MAX_a(x_a) < MIN(\bot_a)$ and so $\bot_a$ set-dominates x_a. If there exists some $x_a \in I_a - \bot_a$ such that $\Gamma_a(x_a, \top_{-a}) > \Gamma_a(\bot_a, \top_{-a}) = \Gamma_a(\bot)$, then $MIN_a(x_a) > MAX(\bot_a)$ and so x_a set-dominates $\bot_a$. Thus, we must have $\Gamma_a(x_a, \top_{-a}) \leq \Gamma_a(\bot_a, \top_{-a}) = \Gamma_a(\bot)$ and $\Gamma_a(x_a, \bot_{-a}) \geq \Gamma_a(\bot)$ for all $x_a \in I_a - \bot_a$. Consequently, $BR_a(\top_{-a}) = \bot_a$ and $BR_a(\bot_{-a}) \neq \bot_a$. This contradicts the seriality of the function BR_a.$\Box$

References

[1] B. D. Bernheim. Rationalizable strategic behavior. *Econometrica*, 52:1007–1028, 1984.

[2] M.S. Chrystall and P. Mars. Adaptive routing in computer communication networks using learning automata. In *Proc. IEEE Nat. Telecomm. Conf.*, pages 121–8, 1981.

[3] D. Fudenberg and D. Kreps. A theory of learning, experimentation, and equilibrium in games. mimeo, Stanford Graduate School of Business, 1988.

[4] D. Fudenberg and J. Tirole. *Game Theory.* The MIT Press, Cambridge, Massachusetts, 1991.

[5] J. Goodman, A. Greenberg, N. Madras, and P. March. Stability of binary exponential backoff. *Journal of the ACM*, 35:579–602, 1988.

[6] F. Gul. Rational strategic behavior and the notion of equilibrium. mimeo, Stanford Graduate School of Business, 1989.

[7] V. Yu. Krylov and M.L. Tsetlin. Games between automata. *Automation and Remote Control*, 24:889–99, 1963.

[8] S. Lakshmivarahan and K. Narendra. Learning algorithms for two-person zero-sum stochastic games with incomplete information. *Mathematics of Oprations Research*, 6:379–86, 1981.

[9] S. Lakshmivarahan and K. Narendra. Learning algorithms for two-person zero-sum stochastic games with incomplete information: a unified approach. *SIAM J. Control and Optimization*, 20:541–2, 1982.

[10] S. Lakshmivarahan and M.A.L. Thatcher. Optimal nonlinear reinforcement schemes for stochastic automata. *Inform. Sci.*, 4:121–8, 1982.

[11] L.D. Mason and X.D. Gu. Learning automata models for adaptive flow control in packet-switching networks. In K.S. Narendra, editor, *Adaptive and Learning Systems*, pages 213–28. Plennum Press, New York, 1986.

[12] P. Milgrom and J. Roberts. Rationalizability, learning and equilibrium in games with strategic complementarities. *Econometrica*, 58:1255–1278, 1990.

[13] P. Milgrom and J. Roberts. Adaptive and sophisticated learning in repeated normal form games. *Games and Economic Behavior*, 3:82–100, 1991.

[14] H. Moulin and S. Shenker. Serial cost sharing. *Econometrica*, 60:1009–1037, 1992.

[15] K. Narendra and M.A.L. Thatcher. *Learning Automata: an introduction*. Prentice Hall, New Jersey, 1989.

[16] D. Pearce. Rationalizable strategic behavior and the problem of perfection. *Econometrica*, 52:1029–1050, 1984.

[17] S. Shenker. Making greed work in networks: a game-theoretic analysis of gateway service disciplines. mimeo, 1989.

[18] S. Shenker. Efficient network allocations with selfish users. In P. J. B. King, I. Mitrani, and R. J. Pooley, editors, *Performance '90*, pages 279–285. North-Holland, New York, 1990.

[19] S. Shenker. A theoretical analysis of feedback flow control. In *Proc. ACM Sigcomm '90*, pages 156–165, 1990.