Synchronous and Asynchronous Learning by Responsive Learning Automata

Eric J. Friedman
Department of Economics, Rutgers University
New Brunswick, NJ 08903.
friedman@fas-econ.rutgers.edu

Scott Shenker
Xerox PARC
3333 Coyote Hill Road, Palo Alto, CA 94304-1314
shenker@parc.xerox.com

November 25, 1996

Abstract

We consider the ability of economic agents to learn in a decentralized environment in which agents do not know the (stochastic) payoff matrix and can not observe their opponents' actions; they merely know, at each stage of the game, their own action and the resulting payoff. We discuss the requirements for learning in such an environment, and show that a simple probabilistic learning algorithm satisfies two important optimizing properties:

i) When placed in an unknown but eventually stationary random environment, they converge in bounded time, in a sense we make precise, to strategies that maximize average payoff.

ii) They satisfy a monotonicity property (related to the "law of the effect") in which increasing the payoffs for a given strategy increases the probability of that strategy being played in the future.

We then study how groups of such learners interact in a general game. We show that synchronous groups of these learners converge to the serially undominated set. In contrast, asynchronous groups do not necessarily converge to the serially undominated set, highlighting the importance of timing in decentralized learning. However, these asynchronous learners do converge to the serially unoverwhelmed set, which we define. Thus, the serially unoverwhelmed set can be viewed as an appropriate solution concept in decentralized settings.

1 Introduction

We consider learning in a repeated-game environment in which agents have extremely limited information. At each stage of the repeated game, each agent takes an action and then observes her resulting payoff. Agents do not know the payoff matrix, which may be stochastic and changing over time, nor can they directly observe the number of other agents or their actions. Because the agents have no *a priori* information about the game structure, they must *learn* the appropriate strategies by trial and error.[1] To gain insight into the general nature of learning in such environments, we study the behavior of a simple probabilistic learning algorithm: the *responsive learning automaton.*[2] This setting, with its extreme lack of *a priori* and observable information, is representative of many decentralized systems involving shared resources where there is typically little or no knowledge of the game structure, and the actions of other agents are not observable.

Computer networks are one example of such a decentralized system, and learning there inherently involves experimentation. One can consider bandwidth usage from a game-theoretic perspective (see, e.g. Shenker (1990,1995)) where the actions are transmission rates and the resulting payoffs are network delays; these delays depend on the transmission rates of all other agents and on the details of the network infrastructure. Network users typically do not know much, if anything, about the actions and preferences of other agents, or about the nature of the underlying network infrastructure. Thus, the agents in a computer network are completely unaware of the payoff structure and must

[1]The degree of uncertainty about the environment is too complex for any kind of Bayesian or "hyper-rational" learning to be practical. This idea is also contained in Roth and Erev (1995) and Sarin and Vahid (1996) who require that in their models of learning, players base their decisions only on what they observe, and do not construct detailed descriptions of beliefs. We do not model such complexity issues formally; There have been many such proposals in the literature (see, e.g. Rubinstein (1986), Friedman and Oren (1995), Knoblauch (1995)) but no consensus on a reasonable measure of complexity.

[2]Our model is based on ordinary Learning Automata which were developed in biology (Tsetlin (1973)) and psychology as models of learning (Norman (1972)) and have been studied extensively in engineering (Lakshmivarahan (1981), Narendra and Thatcher (1989)). Recently, they have also been used by Arthur (1991) as a model of human learning in experimental economics. Also, Borgers and Sarin (1995) have recently compared their behavior with an evolutionary model of learning.

rely on experimentation to determine the appropriate transmission rate. There is a large literature on such experimentation algorithms, called congestion control algorithms; see Jacobson (1988) or Lefelhocz et al. (1996) for a description of the algorithms used in practice and Shenker (1990) for a more a more theoretical description.

Note that learning in this context differs significantly from the standard game theoretic discussion of learning in repeated games, in which agents play best response to either the previous play (Cournot (1838)), some average of historical play (Carter and Maddock (1984)), or some complex Bayesian updating over priors of either the opponents' strategic behavior (Fudenberg and Kreps (1988), Kalai and Lehrer (1993)) or the structure of the game (Jordan (1991)). Note that the first three types learning require the observability of the opponents actions and the knowledge of the payoff matrix, while the fourth will typically be impractical in the situations in which we are interested. Also, none of these allow for the possibility that the payoff matrix may change over time.[3]

In our context, where agents do not even know their own payoff function, trial and error learning is necessary even if the play of other agents is static.[4] Since such 'experimentation' plays a central role, our discussion naturally combines the incentive issues normally addressed by game theoretic discussions of learning with the stochastic sampling theory pivotal to the theory of decentralized control. This combination of issues is of great practical importance in the sharing of decentralized resources (*e.g.*, computer networks, water and power systems), yet is has received rather little attention in the game-theoretic learning literature.

To better understand learning in such contexts, we address two basic questions: (1) what properties should a *reasonable* learning algorithm have in this context, and (2) how can one describe the set of asymptotic plays when all agents use a reasonable learning

[3]In theory, the set of priors could be enlarged to include the possible changes to the payoff matrix; however, this would enlarge the (already large) set of priors substantially. Such computations would clearly be beyond the ability of any real person, or computer.

[4]Kalai and Lehrer (1995) study learning in a similar setting in the context of Bayesian learning.

algorithm? We address these two questions in turn.[5]

What properties should a reasonable learning algorithm have in this context? One natural requirement is that, when playing in a stationary random environment (*i.e.*, the payoff for a given action is in any period is drawn from a fixed probability distribution[6]), the algorithm eventually learns to play the strategy (or strategies) with the highest average payoff. This clearly is an absolutely minimal requirement and was studied by Hannan (1957) and more recently by Fudenberg and Levine (1995).[7] It is also focal in the decentralized control literature. In fact, we posit a stronger condition, which says that even in a nonstationary environment an action which is always optimal (in the sense of expected payoff) should be learned.

There is also a second natural requirement for learning algorithms. Recall that, in our context, players do not directly observe the payoff function, so if the game has changed it can only be detected through their experimentation. Such changes in environment are quite common in many distributed control situations; in computer networks, every time a new user enters or the link topology changes (often due to crashes) – both common occurrences – the underlying payoff function changes. The agents are not explicitly notified that the payoff structure has changed, and thus the learning algorithm should automatically adapt to changes in the environment.

While many standard models of standard models of learning (such as the one introduced by Kalai and Lehrer (1993) and almost all decentralized control algorithms) can optimize against a stationary environment, they are typically not suitable for changing environments. For example, most Bayesian learning procedures (e.g. Kalai and Lehrer

[5]Our approach has much in common with the experimental psychology literature and its recent application by Roth and Erev (1995) to model experiments on extensive form games.

[6]In the simple model we consider, a stationary environment corresponds to i.i.d. actions by other agents and nature. Clearly in other cases stationarity can be interpreted more generally. For example in computer networks there is often cyclical variations, such as day vs. night, in this case stationarity could be defined with respect to the natural cycles. Similarly, simple Markovian effects could also be incorporated into the definition.

[7]Note that in the Bayesian optimality setting (e.g. Kalai and Lehrer, 1995) such convergence may not occur (with high probability) if the agent's priors are sufficiently skewed.

(1993)) and all 'absolutely expedient' (Narendra and Thatcher (1974)) learning algorithms have the property that after a finite time they may discard strategies forever; at any time there is a nonzero probability that they will never play a particular strategy again. Typically, with high probability, these discarded strategies are not currently the optimal strategy; however, if the environment changes and the payoffs for some of these discarded strategies increases, the learner will not be able to detect and react to this change. Thus we conclude that such algorithms are ill-suited to changing environments and we consider only those that are responsive, in the sense just defined.

We are not aware of any algorithms in the economics literature that satisfy both of the aforementioned requirements. We introduce a class of responsive learning automata (RLA) that guarantee responsiveness by requiring that the probability of playing any action (even suboptimal ones) never drop below a fixed constant. This is analogous to the use of mutations in evolutionary learning theory to escape local minima. (In fact, many evolutionary learning models are similarly responsive.) These responsive learning automata satisfy the most important requirements for learning in the environments we consider, and we can thus entertain the second question.

How can one describe the set of asymptotic plays when all agents use a reasonable learning algorithm? Will such learners converge to a Nash equilibrium, or to a "rationalizable" set, or to something else? Will the set of asymptotic play be adequately characterized by traditional game-theoretic concepts? This question is critical for applying the *mechanism design* paradigm (see, e.g., Palfrey (1995)) to decentralized systems, because the set of asymptotic play – the solution concept – will determine which social choice correspondences can be implemented in such settings.

We follow the reasoning developed in Milgrom and Roberts (1991) to describe the collective asymptotic behavior. We show that for RLAs the answer depends critically on the timing of the game. In synchronous settings the automata converge to the serially undominated set, echoing Milgrom and Roberts' result about "adaptive learning". However, in asynchronous settings, the RLAs do *not* necessarily converge to the serially undominated set; in fact they only converge to the serially unoverwhelmed set, which is a much weaker conclusion.

We believe that this result demonstrates that some commonly accepted conclusions about the collective behavior of reasonable learners, such as the convergence to the serially undominated set, do not apply if the play is asynchronous (and if responsiveness is part of the reasonability requirement). Because decentralized systems are typically asynchronous in nature, and responsiveness is crucial in these settings, this implies that the serially undominated set is not an appropriate solution concept for these systems. Thus, lack of synchrony has important implications for learning as it does for other game theoretic analyses (See e.g. Lagunoff and Matsui (1995).) It also has ramifications for decentralized control.

The paper is organized as follows. In Section 2 we define responsive learning automata and then analyze their behavior in the presence of an eventually stationary random payoff function. We describe the collective behavior of multiple automata playing a general game in Section 3; Section 3.1 examines the behavior of synchronous automata, while Section 3.2 examines asynchronous automata. Finally, in Section 4 we briefly consider the case of games which change over time.

2 Learning Automata

In this section we describe the probabilistic learner used in our analysis. We believe that learner is representative of the type of learner's that are sensible in a decentralized setting. The precise assumptions and functional forms that we assume are meant to simplify the exposition and analysis; however, many reasonable probabilistic learners lead to the same conclusions that we obtain for this example.[8] Our definition of a probabilistic learner is a slight variation on standard Learning Automata (LA) ; these modifications improve the convergence properties in a changing environment.

Consider a discrete-time environment with m possible strategies. Let r_i^t denote the payoff for playing strategy i at time t; this payoff, in general, can be random and depend on the entire history of play up to time t. We assume that $0 \leq r_i^t \leq 1$ for all i, t. The state of an automaton consists of the vector of probabilities $p^t = (p_1^t, p_2^t, \ldots, p_m^t)$ where at time t the automaton picks strategy i with probability p_i^t at random. The learning behavior of such automata is embedded in the rule for updating these probabilities.

A standard updating rule, parameterized by a constant $\alpha > 0$, is given by Narendra and Thatcher (1989). If strategy i is picked at time t, then:

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

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

Assume the r_i^t are chosen from some stationary probability distribution, so that the probability that $r_i^t \leq x$ is given by $F_i(x)$ which is independent of t and the history of the plays. These LAs are ϵ-optimal in that, for any $\epsilon > 0$, there exists some α such that:

$$\lim_{t \to \infty} E(\sum_{i=1}^{m} p_i^t r_i^t) > \max(E[r_i^s]) - \epsilon$$

Note that $E[r_i^s]$ is independent of s so the maximum is taken over all i, and see Narendra and Thatcher (1974) for a review of these results.

Because the ϵ-optimality means that these LAs can achieve an asymptotic expected payoff arbitrarily close to the optimal payoff, this property is often cited as evidence that LAs are appropriate learning algorithms. However, the ϵ-optimality property does not mean that for a given sequence of play the average payoff asymptotes to near-optimality. In fact, for these LAs, with probability one the play eventually converges to

[8] This point is formalized in Friedman and Shenker (1996).

a single repeated strategy, so for the sequence of plays from a given LA, $\lim_{t\to\infty} p_i^t = \delta_{i,k}$ for some strategy k (i.e., the limit is zero unless $i = k$ and then the limit is 1).[9] As α goes to zero the probability of this strategy k being optimal approaches one, but for any α there is the possibility that the sequence of plays converges to a nonoptimal payoff.

It is not clear that this eventual collapse to a single strategy, with a nonzero chance of it not being optimal, is adequate for stationary environments. However, such behavior is clearly inappropriate for situations where the environment is not stationary. These LAs, once having converged to a single strategy, are unable to detect any change in the associated payoffs and thus will produce significantly suboptimal performance if the environment (or the other agents' strategies) changed so that the discarded strategies now yielded the maximal payoffs.[10] If one fixes α and then chooses an eventually stationary environment, one can make the probability of the result being suboptimal be substantial.

To rectify this problem of not being able to effectively respond to changing environments, we define a slight variation of the standard LA, which we call a responsive learning automaton (RLA). Essentially, we require that no strategy ever have probability less than $\alpha/2$ of being played[11], thus each strategy will be played infinitely often. The update rule for this RLA, denoted by RLA_α, 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

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

[9]Also, during the tail of this "collapse", there is a high probability that only strategy k is played.

[10]These issues are also briefly discussed in Narendra and Thatcher (1974).

[11]Note that this is analogous to the need for mutations in evolutionary models, where mutations are necessary to keep the replicator dynamics from converging to 'bad' equilibrium .

where $\alpha < 1$. Note that if all $p_j \geq \alpha$ then the update rule for RLA_α is the same as that for the standard learning automaton[12]. We say a vector p^t is *valid* if $\sum_{i=1}^m p_i^t = 1$ and $p_i^t \geq \alpha/2$ for all i. Note that the updating rules transform valid vectors into valid vectors.

We now describe, and verify, two relevant properties of these RLAs: convergence to optimality, and monotonicity.

2.1 Convergence to optimality

A basic requirement for any learning algorithm should be that it in some sense converges to optimality. However, if we wish to consider changing environments, then it is clearly too difficult to be able to optimize in any nonstationary environment. Thus, since we can't require optimality in all environments, we impose the condition that in certain 'simple' environments the learner can optimize. One reasonable condition is that in any environment which is i.i.d. (*i.e.*, stationary) after some finite time, the learner converges to optimality.

We will show that in such an environment RLAs do indeed optimize. First, we note that while LAs achieve ϵ-optimality in stationary environments by eventually discarding all nonoptimal strategies with high probability, the RLAs converge to optimality in a quite different way. RLAs spend most of their time playing optimal strategies, but occasionally wander off and explore other strategies. The manner in which they do this is embodied in the following definition, which we will use to characterize convergence.[13]

Definition 1 *A discrete time random process x^t parameterized*[14] *by α α-converges to 0 if there exist positive constants α_0, β, b_1, b_2, b_3, and q such that, for any $0 < \alpha < \alpha_0$:*

[12]Nonlinear update rules have also been studied (Narendra and Thatcher (1989)), but would not significantly alter our results.

[13]We note that this definition is perhaps not as sharp as possible for our responsive learning automata. This is because we wish to emphasize that our results for multiple automata are not overly dependent on our specific model of learning.

[14]Formally, for each $\alpha > 0$, x_α^t is a random process defined for $t \in Z_+$.

- $\lim_{T\to\infty}\left(\frac{1}{T}\int_0^T dt P[x^t > \sqrt{\alpha}]\right) < \alpha$

- *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 $E[\tau_r] \geq b_2 e^{b_3/\sqrt{\alpha}}/\alpha$.*

α-Convergence is defined by a rapid (polynomial) collapse to near zero and a very long (exponential) period of remaining near zero before random variations take x^t away from zero. In any average of x^t, the exponential period during which x^t remains near zero will dominate the polynomial period which is how long it takes x^t to approach zero from any initial condition. For example, $\lim_{\alpha\to 0^+}\frac{1}{T}\sum_{t=1}^T (x^t)^p = 0$ for any power $p > 0$.

With this definition, we can state the following result, which is a stronger than the optimizing requirement given above and demonstrates the versatility of RLAs.

We consider some automaton in an environment where the random payoffs r_i^t have distributions that in general may depend on the entire history of play.[15] We assume that there is a time T after which there is a particular strategy whose payoffs are maximal (for all histories of the previous play). This obviously includes both stationary and eventually stationary environments with a single optimal strategy, but is much broader. For example, consider a situation in which the environment never converges but there is a particular strategy which is eventually always optimal. Then even in this changing environment the RLA will learn the best strategy.

Theorem 1 *Consider some set of strategies A and define $p_A^t = \sum_{i\in A} p_i^t$. Assume there exists some $\beta > 1$ and $T > 0$ such that $E[r_i^t|h^t] < (1-\frac{m}{\beta})E[r_j^t|h^t]$ for all $t \geq T$, for all $i \in A$, and for all $j \notin A$, where $h^t = (r^1, r^2, \ldots, r^{t-1})$. Then, p_A^t α-converges to 0 from any valid initial condition.*

Thus, when we say that an RLA converges to optimality against a stationary environment, we mean that the probability of playing nonoptimal strategies α-converges

[15]Note that we make no assumptions about the convergence or stationarity of the stochastic process.

to zero. The mixed strategy it plays in each period typically concentrates most of its probability mass on the optimal strategy, but occasionally this mixed strategy vector wanders off to sample nonoptimal strategies. In fact, since the RLA is irreducible, it will eventually come arbitrarily close to every valid mixed strategy. The definition of α-convergence shows that mixed strategy vectors which produce significantly suboptimal outcomes are extremely rare. It follows as a trivial corollary that RLAs are then also ϵ-optimal. Moreover, against a stationary environment where r_i^t is described by the distributions $F_i(x)$, the result of a single run also converges to within ϵ of the optimal payoff: for any $\epsilon > 0$, there exists some α such that:

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \sum_{i=1}^{m} p_i^t r_i^t > \max(E[r_i^s]) - \epsilon$$

almost surely.

Thus, in any eventually stationary environment every single play sequence of an RLA, the long-run payoff approaches the eventually optimal payoff. This is something that the standard LAs fail to do, since while they are close to optimal when averaged over all play sequences, individual play sequences can be significantly suboptimal. The proof of Theorem 1 is in the appendix.

2.2 Monotonicity: The Law of the Effect

An important behavioral property is known as the "Law of the effect" (Thorndike (1898)) which says that strategies that have led to good outcomes in the past are played more often in the future.[16] This can be understood to be a monotonicity requirement on the learning algorithm, in which increasing the payoffs for certain strategies increases the probability that they will be played in the future. We now show that RLAs obey a stochastic formalization of the "law of the effect".

[16] Roth and Erev (1995) consider the law of the effect to be a fundamental principal in understanding players' behaviors in experimental game theory.

Theorem 2 *Consider an automaton RLA_α playing against an environment with a set of payoffs r_i^t, and the same automaton 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 $\hat{p}_A^t = \sum_{i\in A} \hat{p}_i^t$ stochastically dominates $p_A^t = \sum_{i \in A} p_i^t$ 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,

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

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 $i \notin A$. It is easy to see that over any sample path $\hat{p}_A \geq p_A$. □

3 Convergence in Repeated Games with RLAs

These two properties, convergence to optimality (as represented by Theorem 1) and the Law of the Effect (as represented by Theorem 2) can be seen as two simple 'axioms' of decentralized learning. We now turn to an analysis of the behavior of decentralized learners in a noncooperative game. Our key insight is that the asymptotic behavior depends crucially on whether or not the automata are synchronous or asynchronous. We will describe this distinction using the concepts of dominated and overwhelmed strategies.

In this section we first introduce our general model, and then consider the cases of synchronous and asynchronous automata.

3.1 General Model

Consider a game with n players. Assume that player a has m_a possible strategies $\Sigma_a = \{1, 2, \dots, m_a\}$, and let $\Sigma = \Sigma_1 \times \cdots \times \Sigma_n$. For convenience, we will sometimes use the notation $\Sigma = \Sigma_a \times \Sigma_{-a}$. 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 $\Gamma : \Sigma \to \Re^n$ is the payoff function of the game. 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.[17] 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. The collective asymptotic behavior depends on whether or not the automata update their strategies at the same time. We discuss the synchronous and asynchronous cases separately.

3.2 Synchronous Automata

First, we consider the standard model of repeated games, in which there are well defined periods and players update their strategies every period.

A synchronous automaton is one which updates its strategy at every play of the game; the probabilities p are updated and a new strategy is chosen randomly based on these probabilities every time step. We first consider the situation where one of the players is a synchronous RLA, and nothing is known about the nature of the other players. What can be said about the eventual play of the RLA?

To answer this, we apply the concept of dominated strategies. We say that strategy i dominates strategy j for player a, with respect to some set $S_{-a} \subseteq \Sigma_{-a}$, if

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

[17] Equivalently, we could require that each automaton divides her payoff by the largest payoff that she has observed to date. This would not effect any of our results, subject to some boundedness restrictions on the payoffs.

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. Define $U_a : 2^{\Sigma_{-a}} \rightarrow 2^{\Sigma_{-a}}$ for any $S_{-a} \subseteq \Sigma_{-a}$, as the set of undominated strategies:

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

Let $U = (U_1, \ldots, U_n)$. Note that the definition of dominated strategies involved the entire payoff function Γ_a, and that in the context we are considering the players do not know their payoff matrix. Dominated strategies are, in some sense, inferior to undominated ones, and one might expect that a self-optimizing player who knows the entire game matrix and who is playing against opponents with strategies in the set S_{-a} would always play in $U_a(S_{-a})$. We will show that RLAs, despite not knowing the payoff matrix, can also eliminate dominated strategies.

Theorem 3 *Consider some normalized game Γ and a player a whose strategies are chosen by a synchronous RLA. Now assume that the other players choose strategies from S_{-a} with probability greater than $1 - \delta$ in each period. Then, for δ sufficiently small,*

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

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

Proof: Note that when player a is using a synchronous automaton, the plays s_{-a}^t cannot depend directly on the plays s_a^t, except through the vector p^t. We now show that any dominated strategy α-converges to 0.

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 Theorem 2. Furthermore, if we look at the set $A \equiv i \cup M$, then we can apply Theorem 1 to p_A^t to see that is α-converges to 0. Thus p_A^t must α-converge to zero.□

We now consider the case where all the players are synchronous RLAs, so they each update their strategies at every time step according to the RLA updating rules. Can we say more than that they all eliminate dominated strategies? We will show that the collective asymptotic play is restricted to the set $U^\infty(\Sigma)$ where U^∞ represents the infinite iteration of the set mapping U; the set $U^\infty(\Sigma)$ is called the serially undominated set. Our result is very similar to that of Milgrom and Roberts (1991) (albeit in a very different context) who showed that collections of 'adaptive' learners (that is, learners that eventually eliminate dominated strategies) converge to the serially undominated set.

In studying the collective limit, we must apply the concept of α-convergence to a situation where there are n different α's. We restrict ourselves to the case where these α's satisfy the mild restriction that, as the α's approach zero, we always have $\alpha_{\max}^p < \alpha_{\min}$, for some power p where $\alpha_{\max}$ is the largest α and $\alpha_{\min}$ the smallest.

Theorem 4 *For any group of n $(n > 1)$ synchronous responsive learning automata RLA_α playing a normalized game, and for any $p \geq 1$, for any automaton*

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

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

Intuitively, the proof of this theorem follows from repeatedly applying Theorem 3, showing that play collapses to $U^1(\Sigma)$ and then inductively from $U^i(\Sigma)$ to $U^{i+1}(\Sigma)$ until $U^\infty(\Sigma)$ is reached as in Milgrom and Roberts (1991). However due to the stochasticity of the automata, play can occasionally occur outside of $U^i(\Sigma)$. Thus we need to show that this does not destroy the convergence process. We postpone the technical details of the proof until the next section where we prove a similar result in a more intricate setting; see the proof of Theorem 6.

The set $U^\infty(\Sigma)$, the result of the iterated elimination of strictly dominated strategies, has been well studied (see, e.g. Milgrom and Roberts (1991)). Many important learning models have been shown to converge there. In fact for a very large class of games, in particular 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 $U^\infty(\Sigma)$ as shown in Milgrom and Roberts (1991). In these (and other) important models synchronous learning automata converge to the unique equilibrium.

3.3 Asynchronous Automata

In many games there is no natural time period, and thus we consider a game that is being played continuously in time.[18] 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 which she can change any time and then compute the success of the current utilization level as some average over a certain amount of time. This lack of synchrony is quite common

[18] Milgrom and Roberts (1991) show that their analysis also applies to learning in continuous time. However, they still assume the elimination of dominated strategies. As we shall demonstrate, this assumption is difficult to justify in this setting.

in distributed control systems, where time is continuous and the various elements of the system update their behavior independently. For instance, in computer networks, the updating frequency is typically on the order of the round trip time of packets, and this can vary by several orders of magnitude.[19] Thus, there is no synchronization of behavior as we discussed in the previous section.[20]

We model this asynchrony as having the RLA's average their payoffs over some period of time while keeping their strategy fixed. This averaging process must scale inversely with α so that the averaging process occurs on time scales large enough to produce macroscopic changes in the strategy vectors. Thus, as α approaches zero, the asynchronous RLAs average over an asymptotically infinite period to determine their payoff before updating their p's. In this case it is not clear what is the single 'correct' averaging 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 responsive learning automaton which updates its strategy every T/α 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

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

where $G(t)$ is a cumulative distribution function and $s_a^{t'} = i$ for all $t' \in [t - T/\alpha, t]$.

What happens when one player is an $RLA_{\alpha}^{T,G}$, and we know nothing about the other players? In contrast to the results in synchronous automata behavior, the play of the asynchronous RLA is no longer confined to the undominated set. This is because when the RLA holds its strategy fixed over a period T/α, the other players can respond to

[19]The delay in delivering packets on the same ethernet can be several orders of magnitude less than the delivery delay for packets traversing the transatlantic link.

[20]There are few examples of asynchronous games in the literature, and the importance of asynchrony in the play of the game has mostly gone unnoticed, with the exception of recent work by Lagunoff and Matsui (1995).

this strategy.

To characterize the asymptotic behavior of this asynchronous RLA, we introduce the notion of overwhelmed strategies. We say a strategy i for player a overwhelms another strategy j, with respect to S_{-a} if all the possible payoffs associated with i exceed all those payoffs for j:

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

Define $O_a(S_{-a})$ to be the set of unoverwhelmed strategies for player a, if all the other players are playing from the set $S_{-a} \subseteq \Sigma_{-a}$. Unoverwhelmed strategies are a superset of ordinary undominated strategies, $U_a(S_{-a}) \subseteq O_a(S_{-a})$, since the elimination criteria is strictly stronger (i.e., an overwhelmed strategy must also be a dominated strategy, but the converse need not hold).

Overwhelmed strategies, as opposed to dominated strategies, is the appropriate concept when considering asynchronous automata. Even if we assume 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 overwhelms j, because the ordering of the payoffs is independent of how the other players react.

We now show that in asynchronous settings, while RLAs may play dominated strategies, they will not play overwhelmed strategies.

Theorem 5 *Consider some normalized game Γ and a player a whose strategies are chosen by an asynchronous RLA $RLA_{\alpha}^{T,G}$. Now assume that the other players choose strategies from S_{-a} with probability greater than $1 - \delta$ in each period. Then, for δ sufficiently small,*

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

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

Proof: Note that in the asynchronous case the plays s_{-a}^t are not necessarily 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 overwhelmed strategies, U the set of overwhelming 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 Theorem 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.□

As in the synchronous case, we can use this result to characterize the asymptotic collective behavior of a set of asynchronous RLAs repeatedly playing a general game.

Theorem 6 *For any group of n ($n > 1$) asynchronous responsive learning automata RLA_α playing a normalized game, and for any $p \geq 1$, for any automaton*

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

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

Proof: Theorem 6 follows from the repeated application of Theorem 5. For example initially Theorem 5 requires that all players collapse down to the undominated set $S^1 = O(\Sigma)$. Then as all players are in S^1, Theorem 5 now implies that they will collapse down to $S^2 = O^2(\Sigma)$. This process continues until they are all in $S^\infty = O^\infty(\Sigma)$. The same proof applies to Theorem 4 where we replace O by U, Theorem 5 by Theorem 3, and note that all time intervals are the same.

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

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

Now consider $S^i = O^i(\Sigma)$ and let δ_0 be the smallest δ in Theorem 5 for any RLA, a, playing against S^i_{-a}. Similarly define ϕ_1 to be the largest of the quantities $b_1 T/\alpha$, ϕ_2 to be the smallest $b_2 T/\alpha^2$, and ϕ_3 the largest b_3.

Claim 1

$$Pr[\tau_f \leq N \frac{\phi_1}{\gamma \alpha_{min}^2}] \geq 1/2$$

Proof: Choose $\alpha_0 \leq \delta_0$ such that for all $\alpha \leq \alpha_0$ satisfying the restriction $\alpha_{\max}^p \leq \alpha_{\min}$ the following holds,

$$\phi_2 e^{\phi_3/\sqrt{\alpha_{\max}}}/\alpha_{max} > N \frac{\phi_1}{\gamma \alpha_{\min}^2}$$

Now, the above construction guarantees that dominated strategies will never get large during the N repeated actions of the domination operator. Thus an iteration of the domination operator will occur properly.

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

Claim 2

$$E[\tau_f] \leq N \frac{\phi_1}{\gamma \alpha_{\min}^2}$$

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. $\Box$

The result may not provide very much information about the asymptotic play since for many important games $O^\infty(\Sigma)$ 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. This, we believe, is an unavoidable difficulty of learning in asynchronous decentralized settings.

For example, one possible outcome is a Stackelberg equilibrium. Note that Stackelberg equilibria are not possible outcomes of most standard models of learning as shown in Milgrom and Roberts (1991), but it does arise in our model of asynchronous automata. Consider a two automata game when the first automaton (A1) is updating much more often than the second (A2). Then since A1 rapidly converges to the best reply to A2's strategy, we see that A2 is the Stackelberg leader, and they will converge to the Stackelberg equilibrium. More precisely:

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.*[21]

Proof: Choose $T_1 = T_2 = 1$ and $\alpha_2 = \alpha_1^2$. Set $G_1(t)$ be 0 for $t < 1$ and 1 for $t = 1$, while $G_2(t) = t$. Thus player 1 evaluates his payoff at the end of his waiting period while player 2 averages hers over the entire period.

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

Consider their behavior as $\alpha_1 \to 0$. Player 2 updates her strategy $1/\alpha_1$ times while player 1 s strategy is constant. Thus if player 1 is playing $s_1 \in \Sigma_1$, then player 2 will converge to $BR(s_1)$, the Stackelberg follower. Therefore for small enough α_1, player 1 is effectively playing the game $\Gamma_1(s_1, BR(s_1))$, and converges to the Stackelberg leader. □

This is interesting since typically players would prefer to be the leader than a follower. Thus A1 does worse by updating more often than by updating very rarely. This seems counter-intuitive, as one would expect that a rapid response would be a desirable attribute. Thus, using a long averaging interval can be seen as a manipulation of other players.

However, there are certain games where the asynchronous outcome is unambiguous, in that $O^\infty(\Sigma)$ is a singleton. Any set of asynchronous automata will converge to a unique strategy and so no Stackelberg manipulation of the sort described above can occur. One example is the serial cost-sharing game (Moulin and Shenker (1992)). Another class of games which are solvable in terms of unoverwhelmed strategies are those arising in queueing games with many players (Stidham (1992), Friedman and Landsberg (1993)). It is shown in Friedman (1995) that the game is solvable in unoverwhelmed strategies if there is sufficient capacity in the queue. This result extends to many nonatomic games with negative externalities.

4 Time Varying Games

Lastly, we note that our results hold even when the game matrix is not fixed. For example, if the payoffs in each period are random variables with well defined means, then all of our results hold when we define dominated strategies (resp. overwhelmed strategies) in terms of the matrix of expected values. In general the standard models of learning are not so accommodating of stochastic payoffs.

A second interesting extension is when the payoff matrix varies in some systematic manner over time. From our analysis, it seems clear that subject to some regularity assumptions, responsive learning automata will play most of the time in the (current) serially undominated set in the synchronous setting. We now give a simple example of this.

Consider an environment which has a finite number of payoff matrices which are all solvable in terms of dominated strategies, but may have different Nash equilibria. Every τ periods it switches payoff matrices, either randomly or according to a fixed order.

Assume that this game is played synchronously by a group of responsive learning automata. Let $P_\alpha(\tau)$ be the portion of time that the automata are at the (current) Nash equilibrium.

Theorem 8 *Given the above assumptions,* $\lim_{\alpha\to 0}\lim_{\tau\to\infty} P_\alpha(\tau) = 1$ *where the outer limit is taken as all the* α*'s go to zero subject to the restriction that there exists a* $p > 0$ *such that* $\alpha_{\max}^p < \alpha_{\min}$.

Proof: This follows immediately from Theorem 4 and the definition of α-convergence. □

Thus in an environment that changes sufficiently slowly, we still get convergence. Note that for small enough α, we can explicitly compute lower bounds on the percentage of time that play is at Nash. Also, the analogous result holds for asynchronous play.

5 Discussion

In this section we compare and contrast our approach with the literature on learning. This literature began with Cournot's (1838) dynamic interpretation of equilibria. The concept of Nash equilibria did not exist at that time, and so this approach started with a dynamic intended to model reality. It was noticed that this dynamic produced an unambiguous prediction of equilibrium in certain classes of games (those that are best-reply solvable). **is this true, that they noticed?** However, as equilibrium analysis gained prominence, the focus changed from modeling actual dynamics to understanding what dynamic procedures converged to Nash (and other) equilibria.

For instance, in general equilibrium theory, beginning with Walras (1874), much effort went into the study of dynamical mechanisms that converged to the competitive equilibrium. In game theory, a large stream of research was motivated by Robinson's (1951) analysis of fictitious play; many of the subsequent research was devoted to finding dynamic justifications of Nash or other solution concepts.[22] Subsequently, Kalai and Lehrer (1993) showed that Bayesian learning leads to Nash equilibria, subject to a "common priors" assumption. More recently, Foster and Vohra (1996) showed that any "calibrated learner" converges to a correlated equilibrium, where calibration is a natural property of Bayesian learning. Thus correlated equilibria are a 'natural' outcome of Bayesian learning. A number of papers have recently proposed learning algorithms which are guaranteed to be calibrated. (See Foster and Vohra for a review of these calibrated methods, and Fudenberg and Levine (1996) for an overview of fictitious play.)

While Bayesian learning algorithms, and other calibrated algorithms, may be applicable in some settings and are unquestionably important to understanding the foundations of Nash (and correlated) equilibria, we do not believe that these "highly rational" learning algorithms are necessarily applicable in other settings. In particular, in many distributed settings, there is little information about the state of the world (not just the exact nature of the payoff function, but even the number of other players, is unknown). In such "low information" settings, Bayesian learning seems of little practical relevance.[23] More particularly, in computer systems (such as network adaptation al-

[22]Although, to be accurate, Robinson proposed fictitious play as an algorithm for finding Nash equilibria, and the dynamic interpretation came later.**(ck this???)**

[23]Although one could imagine constructing a set of priors over, the number of players, the stochastic environment, the other players' payoffs, etc., this set would obviously be gigantic. However, in order to guarantee convergence using the analysis of Kalai and Lehrer (1993), the prior beliefs over this set would have to include "a grain of truth" – the true state of the world would have to have a nonzero prior. This implies that the priors would have to be nonzero over all potential states of the world, and this computations would require Bayesian updating over this entire support, which would most likely require more computation than is possible using any real computation device.

gorithms) much simpler learning algorithms are used in practice. Similar results are suggested by laboratory experiments.

Our goal in this paper is not to understand the foundations of equilibrium concepts, but to begin a study of learning in a certain nonstandard but important setting. Studying the convergence of a typical learning algorithm leads us to new solution concepts; while these solution concepts may not be terribly appealing from a theoretical viewpoint, they are likely to more accurately represent reality in these "low information/rationality" settings.[24] If one is interested in using mechanism design, or implementation theory, in such settings, one must pay close attention to the nature of equilibrium that results from the learning algorithms used in practice.

Thus, in this paper, we assume that learners make a reasonable, but not necessarily optimal, decisions in the sense described in Theorems 1 and 2. Our work is very close in spirit to that of Roth and Erev (1995) and Arthur (1991). The one crucial difference is that our algorithms embody the notion of "responsiveness" – adapting to changes in the environment within a bounded time (on average) – and this is not built into their models. This responsiveness requirement has significant implications for the convergence behavior of the learning algorithms.

Our results show that these reasonable but relatively naive learners, when compared to more 'rational' learners, have much more difficulty converging to equilibrium in asynchronous environments. In particular, in such settings the correct solution concept is certainly larger than the serially undominated set. We have shown (Theorem 7) that even in games which are dominance solvable (where the convergence of Bayesian algorithms, or even best reply algorithms, are not in doubt), the players may not converge to the Nash equilibrium.

[24] In Friedman and Shenker (1996) ??this forces us to finish that paper!!!?? we show that essentially any group of learners which satisfy Theorems 1,2, and 8, converge to serially unoverwhelmed set, and can remain outside of the serially undominated set by a construction similar to that in Theorem 7.

A Proof of Theorem 1

We prove this theorem with the following sequence of claims. Our approach is to first establish that sufficiently far away from the boundaries p_A^t is decreasing on average and then show that this implies that p_A^t rapidly collapses to near zero as required by the definition of α-convergence. We then bound the escape time from this region near zero. Finally, we show that these results imply the first condition in the definition of α-convergence is obeyed.

Before starting with the claims, note that convergence is independent of the initial valid vector p^0, and thus we can set $T = 1$ with no loss in generality. Choose some $k > 2\beta$. Define $\overline{r}_A^t = \max_{i\in A} E[r_i^t]$ and $\overline{r}_{-A}^t = \min_{i\notin A} E[r_i^t]$. First we show that away from the boundary p_A^t is decreasing on average.

Claim 3 *There exists some constant c_1 such that for $\alpha > 0$ sufficiently small, if $p_A^t > \beta\alpha$ then*

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

Proof: Define $\Delta = E[p_A^{t+1}|p_A^t] - p_A^t$. Computing directly from the updating equations

$$\Delta = \alpha \sum_{j\notin A}\sum_{i\in A}\left\{p_i^t E[p_j^t r_i^t min[1, \frac{p_j^t - \alpha/2}{\alpha p_j^t r_i^t}]] - p_j^t E[r_j^t min[p_i^t, \frac{p_i^t - \alpha/2}{\alpha r_j^t}]]\right\}$$

Noting that $min[1, \frac{p_j^t-\alpha/2}{\alpha p_j^t r_i^t}] \leq 1$ and $min[p_i^t, \frac{p_i^t-\alpha/2}{\alpha r_j^t}] \geq min[p_i^t, \frac{p_i^t-\alpha/2}{\alpha}]$ yields

$$\Delta \leq \alpha \sum_{j\notin A}\sum_{i\in A}\left\{p_i^t p_j^t E[r_i^t] - p_j^t E[r_j^t] min[p_i^t, \frac{p_i^t - \alpha/2}{\alpha}]\right\}.$$

Since $E[r_j^t] \geq \overline{r}_{-A}^t > \overline{r}_A^t(1 - m/\beta)$ for $j \notin A$ and $E[r_i^t] \leq \overline{r}_A^t$ for $i \in A$,

$$\Delta < \alpha\left\{p_A^t p_{-A}^t \overline{r}_A^t (1 - m/\beta) - p_{-A}^t \overline{r}_A^t \sum_{i\in A} min[p_i^t, \frac{p_i^t - \alpha/2}{\alpha}]\right\}.$$

Noting that $\sum_{i\in A} min[p_i^t, \frac{p_i^t-\alpha/2}{\alpha}] \geq p_A^t - m\alpha/2$ shows that

$$\Delta < \alpha p_A^t p_{-A}^t \overline{r}_A^t \left\{(1 - m/\beta) - (1 - \frac{m\alpha}{2p_A^t})\right\}.$$

Since $p_A^t > \beta\alpha$ by assumption, $\Delta < -\alpha^2 \overline{r}_A^t m$, where we again used the fact that $p_A^t > \beta\alpha$ and $p_{-A}^t = 1 - p_A^t$, proving the lemma.$\diamond$

Now we show that p_A^t collapses rapidly to near zero.

Claim 4 *Let τ_f be the first time that $p_A^t \leq \beta\alpha$. Then*

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

Proof: This proof follows that in Goodman et al. (1990). Define

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

This is a submartingale since

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

so

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

and thus

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

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

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

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,

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

completing the proof. $\diamond$

The next two claims show that once $p_A^t \leq 2\beta\alpha$ then it is much more likely go below $\beta\alpha$ before it goes above $k\alpha$. We apply a standard technique from the analysis of submartingales (see e.g. Ross (1983)). We let $\hat{p}_A^t$ denote the process which is the stopped version of p_A^t where stopping occurs as soon as $p_A^t < \beta\alpha$ or $p_A^t > k\alpha$.

Claim 5 *There exists some $c_3 > 0$ such that $e^{c_3 \hat{p}_A^t/\alpha}$ is a submartingale.*

Proof: Let

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

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

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

requires that

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

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

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

Let, $f(c) = E[e^{c(p_A^{t+1} - p_A^t)/\alpha}]$, and note that for $\beta\alpha \leq \hat{p}_A^t \leq k\alpha$, $f(0) = 1$ and $f'(0) = E[p_A^{t+1} - p_A^t] \leq -c_1\alpha^2$. Therefore there must exist some constant $c_3 > 0$ such that $f(c_3) < 1$. For this constant, z^t is a submartingale.$\diamond$

Claim 6 *If $\hat{p}_A^0 \leq 2\beta\alpha$ then there exists some constant c_4 such that*

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

Proof: Let P_k^t be the probability that $\hat{p}_A^t > k\alpha$, P_f^t be the probability that $\hat{p}_A^t < \beta\alpha$, and $z^t = e^{c_3\hat{p}_A^t/\alpha}$ For all $t > 0$ we have

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

Note that p_A^t is ergodic while in $\hat{p}_A^t$ the states $z^t < e^{\beta c_3}m$ and $z^t > e^{kc_3}$ are absorbing states. Thus $P_f = \lim_{t\to\infty} P_f$ and $P_k = \lim_{t\to\infty} P_k$ both exist (since the sequences are monotone) and sum to 1.

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

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

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

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

We can use the two preceding claims to show that the 'escape' time is large. This is done in the next two claims.

Claim 7 *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*

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

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

$$E[\tau_k] \geq E[\text{number of times to } \alpha \text{ before } k\alpha]/2\alpha = \frac{1}{2\alpha P_k} > \frac{e^{c_3 k}}{2 c_4 \alpha} \qquad \diamond$$

Claim 8 *Assume that $p_A^0 \leq 2\beta\alpha$. Let τ_r be the first time that $p_A^t \geq \sqrt{\alpha}$. Then,*

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

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

Since collapse is polynomially fast and escape exponentially slow, it is easy to see that the limiting probability density is concentrated near 0. Thus, we can show that the first condition in the definition of α-convergence is obeyed.

Claim 9 *There exists some α_0 such that for all $\alpha < \alpha_0$*

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

Proof: 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 occurred more recently than $p_A^{t'} > \sqrt{\alpha}$, and C:$p_A^t \geq \beta\alpha$ and $p_A^{t'} > \sqrt{\alpha}$ has occurred 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.$\Box$

Note that we have also established the rate of convergence for small but finite α. (For example, if $0 < \alpha < \alpha_0$ then p_A will become smaller than $\sqrt{\alpha}$ in time proportional to $1/\alpha^2$ where the exact time is given in the above lemmas.)

References

[1] W. B. Arthur. Designing economic agents that act like human agents: A behavioral approach to bounded rationality. *Learning and Adaptive Economic Behavior*, 81(2):353–9, 1991.

[2] T. Borgers and R. Sarin. Learning through reinforcement and replicator dynamics. Mimeo, 1996.

[3] M. Carter and R. Maddock. *Rational Expectations.* MacMilan, London, 1984.

[4] A. Cournot. *Recherches sur les Principes Mathematics de la Theorie de la Richesse.* Hachette, Paris, 1838.

[5] D. Foster and R. Vohra. Calibrated learning and correlated equilibrium. Mimeo, 1996.

[6] E. J. Friedman. Non-atomic games with multiple externalities. mimeo. Durham North Carolina: Duke University., 1995.

[7] E. J. Friedman and A. S. Landsberg. Short run dynamics of multi-class queues. *Operations Research Letters*, 14:221–229, 1993.

[8] E. J. Friedman and S. S. Oren. The complexity of resource allocation and price mechanisms under bounded rationality. *Economic Theory*, 6:225–250, 1995.

[9] E. J. Friedman and S. Shenker. Decentralized learning and the design of the internet. mimeo., 1996.

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

[11] D. Fudenberg and D. Levine. Consistency and cautious fititious play. *Journal of Economic Dynamics and Control*, 19:1065–1089, 1995.

[12] D. Fudenberg and D. Levine. Theory of learning in games. Mimeo, 1996.

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

[14] J. Hannan. Approximation to bayes risk in repeated play. In M Dresher, A. Tucker, and P. Wolfe, editors, *Contributions to the Theory of Games*, volume 3, pages 97–140. Princeton University Press, Princeton, NJ, 1956.

[15] V. Jacobson. Congestion avoidance and control. In *Proc. ACM Sigcomm '88*, pages 314–329, 1988.

[16] J. Jordan. Bayesian learning in normal form games. *Games and Economic Behavior*, 3:60– 81, 1991.

[17] E. Kalai and E. Lehrer. Rational learning leads to nash equilibria. *Econometrica*, 61(5):1019–1045, 1993.

[18] E. Kalai and E. Lehrer. Subjective games and equilibria. *Games and Economic Behavior*, 8:123–163, 1995.

[19] M. Kandori, G. Mailath, and R. Rob. Learning, mutation, and long run equilibria in games. *Econometrica*, 61:29–56, 1993.

[20] V. Knoblauch. Computable strategies for the repeated prisoner's dilemma. *Games and Economic Behavior*, 7(3):381–89, 1994.

[21] R. Lagunoff and A. Matsui. An "anti-folk theorem" for a class of asynchronously repeated coodination games. Mimeo, 1995.

[22] S. Lakshmivarahan. *Learning Algorithms: theory and applications*. Springer-Verlag, New York, 1981.

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

[24] C. Lefelhocz, B. Lyles, S. Shenker, and L. Zhang. Congestion control for best effort service: why we need a new paradigm. *IEEE Network*, 10:10–19, 1996.

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

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

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

[28] K. Narendra and M.A.L. Thatcher. Learning automata: a survey. *IEEE transactions on systems, man, and cybernetics*, SMC-4(4):889–99, 1974.

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

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

[31] T. Palfrey. Implementation theory. mimeo:Caltech, 1995.

[32] J. Robinson. An iterative method of solving a game. *Annals of Mathematics*, 54:296–301, 1951.

[33] S. Ross. *Stochastic Processes*. John Wiley and Sons, New York, 1983.

[34] A. Roth and I. Erev. Learning in extensive form games: experimental data and simple dynamic models in the intermediate term. *Games and Economic Behavior*, 8:164–212, 1995.

[35] A. Rubinstein. Finite automata play the repeated prisoners dilemma. *Journal of Economic Theory*, 39:83–96, 1986.

[36] R. Sarin and F. Vahid. A simple model of dynamic choice. Mimeo, 1996.

[37] 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.

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

[39] S. Shenker. Making greed work in networks: A game-theoretic analysis of switch service disciplines. *IEEE/ACM Transactions on Networking*, 3:819–831, 1995.

[40] J. Maynard Smith. *Evolution and the theory of games*. Cambridge University Press, Cambridge, 1982.

[41] S. Stidham. Pricing and capacity decisions for a service facility: Stability and multiple local optima. *Management Science*, 38(8):1121–1139, 1992.

[42] E.L. Thorndike. Animal intelligence: an experimental study of the associative processes in animals. *Psychol. Monogr.*, 2, 1898.

[43] M.L. Tsetlin. *Automaton Theory and Modelling of Biological systems*. Academic Press, New York, 1973.

[44] L. Walras. *Elements d'Economie Politique Pure*. Corbaz, Lausanne, 1874.