Reinforcement Learning in a Nonstationary Environment: The *El Farol* Problem

Ann Maria Bell[1]

September 8, 1999

[1]Caelum Research, NASA Ames Research Center, Mail Stop 269-3, Moffett Field, CA 94035-1000, abell@mail.arc.nasa.gov.

Abstract

This paper examines the performance of simple learning rules in a complex adaptive system based on a coordination problem modeled on the *El Farol* problem. The key features of the *El Farol* problem are that it typically involves a medium number of agents and that agents' payoff functions have a discontinuous response to increased congestion. First we consider a single adaptive agent facing a stationary environment. We demonstrate that the simple learning rules proposed by Roth and Er'ev can be extremely sensitive to small changes in the initial conditions and that events early in a simulation can affect the performance of the rule over a relatively long time horizon. In contrast, a reinforcement learning rule based on standard practice in the computer science literature converges rapidly and robustly. The situation is reversed when multiple adaptive agents interact: the RE algorithms often converge rapidly to a stable average aggregate attendance despite the slow and erratic behavior of individual learners, while the CS based learners frequently over-attend in the early and intermediate terms. The symmetric mixed strategy equilibria is unstable: all three learning rules ultimately tend towards pure strategies or stabilize in the medium term at non-equilibrium probabilities of attendance. The brittleness of the algorithms in different contexts emphasize the importance of thorough and thoughtful examination of simulation-based results.

1 Introduction

When small numbers of agents interact we expect that they will behave strategically and anticipate the behavior of other agents. When large numbers of agents interact we often assume that an individual's action has a neglible effect on the behavior of the system and that agents do not engage in strategic behavior. Exploring the intermediate case where the interactions of a medium-sized population of agents create an endogenously evolving environment is one of the goals of complex systems theory.

This paper utilizes the *El Farol* problem proposed by W. Brian Arthur [1] to explore the effects of endogeneity in a complex adaptive system where agents use simple reinforcement learning rules. We begin by examining the performance of three different learning rules when a single learning agent faces a stationary stochastic environment, i. e. when all other agents choose their actions based on independent draws from a fixed probability distribution. We demonstrate that details of the problem specification which do not affect the Nash equilibria of the underlying game such as the initial conditions of the adaptive agents' states and the scaling of rewards can dramatically affect the performance of the learning rules. We then analyze the changes in the dynamic and equilibrium behavior of the system as the proportion of adaptive learning agents increases, creating an endogenously evolving, non-stationary environment. In many cases, the adaptive system as whole rapidly converges to a fixed average or aggregate behavior despite the often slow and erratic convergence of the individual learning rules. But again, the details of the algorithm specification can lead to very different global individual outcomes, especially in the short and medium term.

Reinforcement learning (RL) is a powerful technique studied extensively by computer scientists that focuses on how agents learn from interacting with their environment rather than by forming explicit models of the environment or by searching the policy space through evolution and selection [20]. Agents formulate policies, or mappings from states to actions, on the basis of the rewards associated with those state-action pairs in the past. Reinforcement learning can readily accommodate uncertainty about the environment and delayed

consequences of actions. The theoretical analysis has concentrated, however, on stationary environments. Although RL techniques in computer science focus on learning through interactions with the environment they are closely related to dynamic programming approaches and other planning methods. We contrast the performance of the standard formulation of an RL algorithm in the computer science literature with the performance of two simpler reinforcement learning algorithms utilized by Roth and Er'ev ([17], [9]) which are loosely based on the work of psychologists Bush and Mosteller [5].

The next section motivates and reviews the *El Farol* problem. The following section states the three different RL algorithms. The fourth section uses simulations to illustrate the behavior of the individual learning rules in isolation and of the larger complex adaptive system. The final section concludes.

2 The *El Farol* Problem

El Farol is a bar in Santa Fe. The bar is popular, but becomes overcrowded when more than sixty people attend on any given evening. Everyone enjoys themselves when fewer than sixty people go, but no one has a good time when the bar is overcrowded. In the absence of information about other people's choices how can, or how do, people choose whether to go to *El Farol*?

The *El Farol* problem is considered a canonical example of a complex adaptive system [6]. Arthur originally posed the *El Farol* problem to illustrate the aggregate dynamics of a system composed of bounded rational agents who rely on inductive learning. Agents attempt to predict the aggregate behavior of other agents, which simultaneously depends on all agents' predictions. Consequently, the interaction between individual learning strategies and the environment that agents face plays a key role in determining the dynamics of the system. In contrast to many game theoretic treatments of learning and coordination, the level of congestion at *El Farol* depends on the actions of a relatively large number of individual agents. It emphasizes the difficulty of coordinating the actions of independent agents

without a centralized mechanism and provides a simple scenario for examining congestion and coordination problems that arise in large, rapidly evolving systems like the Internet. These features make *El Farol* a useful tool for analyzing information technology systems which are characterized by decentralized decision making and rapid endogenous changes in the operating environment[1]. The *El Farol* problem has received a significant amount of attention from physicists and computer scientists[2].

We consider the *El Farol* problem as a one-shot simultaneous move game[3]. It is a multi-player congestion game where the payoffs depend only on the number of agents choosing that action. Congestion games were first characterized by Rosenthal [16]. Finding a Nash equilibrium of a congestion game is equivalent to a constrained minimization problem. Congestion games are isomorphic to potential games [15].

Let all agents have identical payoffs: b is the payoff an agent receives for attending a crowded bar and g is the payoff an agent receives for attending an uncrowded bar. Without loss of generality let h, the payoff received for staying home, be zero. (Some of the algorithms considered below require weakly positive rewards.) Let M be the total number of agents and $\mathcal{N}$ be the maximum capacity of an uncrowded bar. The game is then $G = [M, \{S_i\}, u_i(s_i, s_{-i})]$ where S_i consists of two strategies, stay home (indicated by 1) and go to the bar (indicated by 2) with payoffs determined by $u_i(0, s_{-i}) = 0$ for all s_{-i},

[1]The analogy between the *El Farol* problem and decentralized resource allocation is discussed by Greenwald et. al. [12], as well as in our previous work [2] [19]. Glance and Huberman [11] and Huberman and Lukose [13] also consider the dynamics of congestion on the Internet in a game theoretic framework.

[2]Johnson, Jarvis, Jonson, Cheung, Kwong and Hui [14] consider how the variance in the *El Farol* problem changes in response to the number of predictors available in the entire system and the number of predictors that each agents selects. Edmonds [8] expands Arthur's approach by endowing the agents with an evolving set of predictors, and by allowing communication between agents (including the ability for agents to 'lie' to each other). Zambrano [23] applies results from Bayesian game theory to show that a system composed of Bayesian learners will converge to the set of Nash equilibria. Wolpert, Tumer and Wheeler [21] and Wolpert and Tumer [22] consider a variant of the *El Farol* scenario and design utility functions and learning algorithms for individual agents that collectively optimize a global welfare function without centralized control. Challet and Zhang [7] simplify the *El Farol* problem by considering a 'minority game' in which receive positive payoffs when they choose to join the smaller of two groups using strategies which consist of a table that maps the outcome in a fixed number of previous periods to a choice of group next period. Savit, Manuca, Li and Riolo [18] utilize the 'minority game' as a simple model of co-evolving adaptive systems.

[3] Portions of this section describing the characteristics of the *El Farol* game follow the exposition in [2].

$u_i(1, s_{-i}) = g$ when $\sum_{-i} \leq \mathcal{N} - 1$, and $u_i(1, s_{-i}) = b$ when $\sum_{-i} > \mathcal{N} - 1$.

In a deterministic setting where agents utilize only pure strategies a Nash equilibrium occurs when exactly sixty agents choose to attend. There are $\binom{100}{60}$ such equilibria. There are no symmetric pure strategy Nash equilibria. Pure strategy Nash equilibria are Pareto efficient. Unlike the standard public good framework, in the *El Farol* scenario fully informed optimizing agents will not increase consumption of a publicly available resource until it experiences an inefficient level of congestion: if agents could predict the behavior of other agents perfectly the bar would never be crowded and all patrons would have a good time[4]. Coordination failure, or agents' uncertainty about the action of other agents, may be an important source of congestion in large decentralized systems [2].

Because the payoffs in the *El Farol* game contain a discontinuous response to increased attendance, the analysis of equilibria depends crucially on how the agent accounts for his or her own behavior. Each agents' mixed-strategy profile consists of a single parameter p_i which indicates the probability that agent i attends. Let M be the total number of agents, N be the total observed attendance, N^{-i} be the observed attendance exclusive of agent i and $\mathcal{N}$ be the maximum capacity of an uncrowded bar.

A mixed-strategy equilibria must satisfy the condition:

$$g\ Pr\left(N^{-i} \leq \mathcal{N} - 1\right) + b\ Pr\left(N^{-i} > \mathcal{N} - 1\right) = 0$$

$$\text{or} \quad Pr\left(N^{-i} \leq \mathcal{N} - 1\right) = \frac{b}{b - g} \tag{1}$$

which states that the expected return to the pure strategy of attending the bar exactly equals the expected return to the pure strategy of staying home. This must hold for all agents simultaneously. Also note that the indifference condition that determines a mixed strategy equilibrium depends on the distribution of total attendance which in general depends on the probabilities of attendance for individual agents, not just on the mean of the entire distribution.

[4]The stochastic or mixed-strategy framework may suffer from socially inefficient congestion as discussed below.

For a symmetric mixed strategy equilibria the probability that $\mathcal{N}-1$ or fewer agents attend is :

$$\sum_{N^{-i}=0}^{N^{-i}=\mathcal{N}-1} \binom{\mathcal{N}-1}{N^{-i}} \left(p^{N^{-i}}(1-p)^{\mathcal{N}-1-N^{-i}}\right). \tag{2}$$

For the case where $M = 100$ and $\mathcal{N} = 60$ this involves finding the roots of a 100th order polynomial. For example, when $g \approx 2.02$, $b = 0$, and $g \approx 1.02$ the symmetric mixed-strategy equilibrium is $p = .6$.

The symmetric mixed strategy Nash equilibrium is not Pareto optimal. Agents should increase their probability of attending unless the expected return to attendance exactly equals that of staying home. The randomness in agents' choice of strategy will generate Pareto inefficient variance in attendance. Any attendance outcome that falls short of the maximum capacity of an uncrowded bar can be Pareto improved by increasing attendance, and vice versa. The Pareto efficient symmetric mixed-strategy profile[5] can be calculated by:

$$\max_{p} \sum_{N=0}^{N=\mathcal{N}} g\ N\ Pr(N) + \sum_{N=\mathcal{N}+1}^{N=M} b\ N\ Pr(N). \tag{3}$$

This p maximizes the total expected payoff to all agents which also maximizes the expected return to individual agents. When $g \approx 2.02$, $b = 0$, and $g \approx 1.02$ as above the Pareto efficient symmetric mixed-strategy profile is $p \approx .5$. In this sense the *El Farol* problem suffers from inefficient congestion similar to that observed in a standard public goods framework: each individual agent's probability of attendance is just high enough that the expected return is the same as staying home.

There are no asymmetric mixed strategy equilibria. Consider two agents with differing probabilities of attendance and, without loss of generality, label them agents 1 and 2 with $p_1 < p_2$. The indifference condition (1) must hold for every agent, which implies that $Pr\,(N^{-1} \leq \mathcal{N}-1)$ equal $Pr\,(N^{-2} \leq \mathcal{N}-1)$. The density function for attendance exclusive of agent 1 can be expressed in terms of the density function for attendance exclusive of agents

[5]The mixed-strategy profile that maximizes the expected return to each agent given the constraint that the expected return be equal for all agents.

1 and 2:

$$Pr(N^{-1} = 0) = Pr(N^{-1,-2} = 0)(1 - p_2)$$

$$Pr(N^{-1} = x) = Pr(N^{-1,-2} = x)(1 - p_2) + Pr(N^{-1,-2} = x - 1)\, p_2.$$

By expanding and combining sums the cumulative distribution that agent 1 faces can be expressed as:

$$Pr(N^{-1} \leq X) = \sum_{x=0}^{x=X-1} Pr(N^{-1,-2} = x) + Pr(N^{-1,-2} = X)(1 - p_2).$$

The cumulative distribution function that agent 2 faces differs only by the term $(1-p_2)$ which is replaced by $(1 - p_1)$. Consequently, the indifference condition cannot hold simultaneously for two agents with different probabilities of attendance.

3 Reinforcement Learning

In the simplest case there is a single learning agent who faces a stationary stochastic environment: all of the other agents choose to attend the bar with a fixed probability. The future behavior of the system is exogenously determined and independent of the agent's actions. The game reduces to a two-armed bandit problem in which one of the actions (staying home) returns a fixed payoff and the other action (attending the bar) returns a payoff determined by the realization of a multinomial random variable corresponding to the actions of the other agents. In this simple scenario the agent's ability to learn the correct action demonstrates the effectiveness of the various learning rules.

The situation becomes more complex when two agents are learning at the same time: the environment is still largely exogenous and stationary but the possibility that one agent's actions will influence the future actions of the other agent is introduced. As a larger fraction of the agents simultaneously attempt to learn the value of the two possible actions the environment becomes less stationary, at least initially, and the future behavior of individual agents and the entire system depends on the dynamic interaction of the learning rules.

We consider an algorithm based on standard practice in the machine learning literature [20]. At time t each learning agent i is specified by a vector of weights for the two actions: $w_t^i = \{w1_t^i, w2_t^i\}$, and a vector which records the number of times that each action has been taken: $b_t^i = \{b1_t^i, b2_t^i\}$. In a basic reinforcement learning algorithm for reward maximization agent i chooses each action with probabilities:

$$p1_t^i = \frac{e^{\frac{w1_t^i}{T_t}}}{e^{\frac{w1_t^i}{T_t}} + e^{\frac{w2_t^i}{T_t}}} \qquad \text{and} \qquad p2_t^i = \frac{e^{\frac{w2_t^i}{T_t}}}{e^{\frac{w1_t^i}{T_t}} + e^{\frac{w2_t^i}{T_t}}} \tag{4}$$

where T_t is a "temperature" parameter that declines slowly over time, making it more likely that the action with the higher weight is chosen. The function determining T_t is:

$$T_{t+1} = Max[\mu\, T_t, \bar{T}] \tag{5}$$

where $\bar{T}$ is a parameter indicating the minimum temperature and μ, $0 < \mu < 1$, is a multiplicative stepsize. The initial value of T_0 is a parameter that indicates the amount of randomness in agents' choices early in the simulation. Whenever $T_t = 0$ the agent deterministically chooses the action with the higher weight.

The weights for an action are updated according to the following rules:

$$\begin{aligned} w1_{t+1}^i &= w1_t^i + I1_t^i\, \beta 1_t\, (-w1_t^i + r1_t) \\ w2_{t+1}^i &= w2_t^i + I2_t^i\, \beta 2_t\, (-w2_t^i + r2_t) \end{aligned} \tag{6}$$

where $I1_t^i$ and $I2_t^i$ are indicator variables that equal one when the action is taken and zero when it is not, and $r1_t$ and $r2_t$ are the rewards at time t for taking actions 1 (staying home) and 2 (going to the bar) respectively. The parameters $\beta 1_t = \frac{1}{b1_t^i}$ and $\beta 2_t = \frac{1}{b2_t^i}$ for updating the weights correspond to averaging the payoff for that action over all observations. Consequently, the weights change more slowly over time. In a non-stationary environment it is more appropriate to have $\beta 1$ and $\beta 2$ constant to accommodate the changing rewards for actions. The theoretical results from the literature and simulations demonstrate the specification of $\beta 1$ and $\beta 2$ can affect the performance of the algorithm significantly. Note

that the rewards may be negative: the probabilities in (4) above are defined for negative and positive weights. We refer to the algorithm determined by (4), (5) and (6) and the β vector as the CS algorithm.

The key free parameters are T_0, μ and $\bar{T}$. The final temperature $\bar{T}$ determines how much experimentation the agents engage in in the long run but does not influence the algorithm before that point. When $\bar{T}$ is zero agents eventually utilize a pure strategy. The initial temperature and the stepsize play a crucial role in determining the speed of convergence: a low initial temperature or one that decreases too quickly can lead to slow convergence because the agents are unlikely to update the weights for initially less preferred actions; a high initial temperature or one that decreases too slowly can lead agents to choose actions without much regard to the average payoffs of those actions. The initial weights $\{w1_0^i, w2_0^i\}$ are also free parameters, but for reasonable initial values of T_0 they have little effect on the behavior of the system: the differences in weights are relatively unimportant when T is high, and the initial weights are replaced with the reward received after the first time an action is chosen.

Roth and Er'ev [17] and Er'ev and Roth [9] consider reinforcement learning rules inspired by results in psychology. They also use a vector of weights for each action but the weights are translated into probabilities by:

$$p1_t^i = \frac{w1_t^i}{w1_t^i + w2_t^i} \qquad p2_t^i = \frac{w1_t^i}{w1_t^i + w2_t^i}. \tag{7}$$

They consider several different ways to update the the weights, two of which are considered here. In the first case the rewards are added to the weights for each action:

$$\begin{aligned} w1_{t+1}^i &= w1_t^i + I1_t^i \, r1_t \\ w2_{t+1}^i &= w2_t^i + I2_t^i \, r2_t \end{aligned} \tag{8}$$

The weights are the total sum of rewards for each action. We refer to (7) and (8) along with the initial conditions of the parameters as the RE1 learning rule[6]. The free parameters are

[6]This is referred to as "the basic reinforcement model" in [9], p. 860. Because the minimum payoff is

the two initial weights $\{w1_0^i, w2_0^i\}$, which determine the initial probability and the scaling of the stepsize or the extent to which the initial rewards received change the probability. The magnitude of the rewards also affects the rate of change of the probabilities: the algorithm is not independent of the units of measurement. Note that every time an action is taken it is more likely to be taken in the future: high rewards received early in the simulation can have a large affect on the future trajectory of the system. Roth and Er'ev address this by including a "forgetting" parameter in the updating of the weights:

$$w1_{t+1}^i = (1-\phi)w1_t^i + I1_t^i\, r1_t$$
$$w2_{t+1}^i = (1-\phi)w2_t^i + I2_t^i\, r2_t \tag{9}$$

where ϕ prevents the weights from growing without bound over time and puts a lower bound on the change in the probability of taking an action for non-zero rewards. We refer to (9) and (7) and the initial parameters $\{w1_0^i, w2_0^i\}$ and ϕ as the RE2 rule[7].

The next section demonstrates the behavior of these algorithms in the context ofthe *El Farol* problem.

4 Simulation Results

4.1 A single learner in a stationary environment

The initial simulations explore the behavior of the different algorithm specifications when an individual learner faces a stationary environment. The reward for attending an uncrowded bar is ≈ 2.02, for staying home, ≈ 1.02, and for attending a crowded bar, 0. There are 30 agents, 29 of whom base their actions on independent realizations of a Bernoulli random variable with probability of attending of .61. The expected payoff to attending the bar is ≈ 0.94. Consequently, the best response of the learning agent is to always stay home.

zero, hence the term they refer to as x_{min} disappears.

[7]This rule is stated on p. 863 in [9] and on p. 175 in [17].

The initial parameters for the algorithms determine the probability of attending in the first period, this is .60 for all three algorithm specifications. For the CS algorithm the initial weights are $w_0 = \{1.02, 1.32\}$ which leads to a 60% chance of attending when combined with the initial temperature $T_0 = .75$. The multiplicative factor used to lower the temperature over time is $\mu = .9975$ until the minimum temperature of $\bar{T} = .025$ is reached. For the RE1 and RE2 algorithms the initials weights are $w_0 = \{.8, 1.2\}$ and the forgetting parameter is $\phi = .001$.

Figure 1 shows the probability of attendance for the learning agent for 50000 iterations of all three algorithms. The same random numbers were used to determine the action taken and the attendance of other agents in all three cases[8]. This suggests the long run behavior of the algorithms. The top line in the figure is the RE1 algorithm, the middle line is RE2 and the bottom line is CS. The CS algorithm rapidly trends down to and then fluctuates around the probability of attendance associated with the correct estimate of the value of the two actions and the minimum temperature of .025. The RE2 algorithm continues to decline over time. The RE1 algorithm, although apparently stuck at a high probability of attendance, also declines over time: the expected change in the probability of attendance is negative at every time step[9].

Figure 2 is a close up of Figure 1 which shows the probability of attending for 5000 iterations. In this time frame the probability of attendance for the RE2 algorithm is slightly

[8]The pseudo-random number generator chooses a real number between zero and one. If that number is below the probability of attending then the outcome of the Bernoulli random variable is one (attend), zero (stay home) otherwise. Consequently, when the probabilities of attending are similar the same action is likely to be returned. The pseudo-random number generator acts as an external signal: the differences in the behavior of the algorithms arises from different responses to the same signal. Also, the realizations of the multi-nomial random variable determining the attendance of the non-adaptive agents is the same across simulations. This provides a more accurate basis for comparison of the performance of the algorithms.

[9]The difference between the probability of attending at time t and the expected value of attending at time $t+1$ is:

$$\frac{w2\,((w1+w2)\,(2\,w2\,(w1+w2)+h\,(w1+2\,w2))+g\,((w1+w2)\,(P\,w1+2\,w2)+h\,(w1+P\,w1+2\,w2)))}{(w1+w2)^2\,(g+w1+w2)\,(h+w1+w2)}$$

which is always greater than zero but declines rapidly over time as $w1$ and $w2$ increase. (P is the probability that $c-1$ other agents choose to attend.)

higher than the RE1 algorithm; the CS algorithm is still the lower line. The RE algorithms both move in the wrong direction, increasing from the initial probability of attending of .60 to a maximum value of $\approx$ .80. Figure 2 demonstrates the behavior of algorithms in the "intermediate term" that Roth and Er'ev loosely define as the time it takes for the learning curve to become very flat. Comparing the two figures shows the difficulties that arise in identifying the intermediate term. The apparent stability of the RE2 algorithm's probability of attendance disappears after the first 5000 iterations even though the stepsize (magnitude of the change in probability) continues to decline over time.

One of the key ingredients in these (and all) adaptive algorithm is the "stepsize" or the amount the probability of attending changes at each time step, which is influenced by several parameters in these algorithms. Figures 3, 4 and 5 show the change in the probability of attending over time for the three algorithms. Early in the simulations the magnitude of the changes in the probability of attending are roughly equal in all three cases[10]. The sum of the absolute values of the change in probability over the first 25 iterations for the CS algorithm is $\approx$ .036, for the RE algorithms, $\approx$ .037. After 500 iterations it is $\approx$ .005 for CS algorithm and $\approx$ .004 for the RE algorithms. (Much later in the simulation the stepsize for the RE2 algorithm becomes larger than that of the RE1 algorithm.) The differences in their behavior are not explained by differing stepsizes, instead, it results from the way the algorithms incorporate the rewards: the CS algorithm decreases the likelihood of going to the bar after a bad experience (the zero payoff is averaged into the weight for attending) but does not change the likelihood of attending after staying home (the average of the fixed reward doesn't change after the first few times the action is taken); in contrast, the RE algorithms do not change the likelihood of attending after a bad experience (adding a zero leaves the weight for attending unchanged) but decreases the likelihood of attending after staying home (the positive payoff is added into the weight on staying home). The tendency

[10]The first observation of $\approx$ .3 for the CS algorithm is not shown on the graph in order to keep the scale the same in the figures for all three algorithms.

to (weakly) increase the probability of any action that has been taken can lead the RE algorithms away from the optimal action in the short and medium term.

The performance of the RE1 algorithm is more problematic than the previous figures suggest: slight changes in the initial conditions can dramatically alter the observed behavior. Figure 6 shows two simulations of the RE1 algorithm with different initial conditions but the same underlying sequence of random numbers as in figure 1. In the first case (upper solid line) the initial weights are $\{.8, 1.0411\}$ with an initial probability of attendance of ≈ 0.565477; the behavior is similar to that observed previously. In the second case (lower cashed line) the initial weights are $\{.8, 1.04105\}$ with an initial probability of attendance of ≈ 0.565465; here the RE1 algorithm rapidly tends towards staying home, the optimal action. The difference in the initial probabilities of attendance between the two cases is ≈ 0.0000118; the difference in the probability of attending at time 10000 is ≈ 0.72. Figure 7 shows the first 50 iterations of figure 1. The divergence of the two simulations with different initial conditions occurs in the first iterations: these continue to influence the adaptive agents behavior over the entire course of the simulation.

Roth and Er'ev [17] do not consider the initial probability to be one of the free parameters of their model: they set it to 50% in all cases. This assumption can play a crucial role in the behavior of the learning rule. The one free parameter they consider is the scaling or magnitude of the initial weights. Increasing the magnitude of the initial weight makes the changes in probability smaller, especially in the first few time steps. Changing the size of the initial weights in this example changes where the divergence in behavior occurs but does not qualitatively change the result: figure 8 shows two simulations with initial weights of $\{8, 9\}$ (upper solid line) and $\{8, 8.75\}$ (dashed lower line). The larger initial weights slowed the movement of both adaptive learners: the difference in the probability of attending at time 10000 is ≈ 0.25. A similar situation arises with the RE2 algorithm: figure 9 shows the trajectory of the probability of attendance starting from initial weights $\{.8, 1.05\}$ (upper solid line) and $\{.8, 1.04\}$ (lower dashed line). Although the RE2 algorithm eventually declines

overtime, the effects of the initial conditions are apparent after tens of thousands of iterations.

The previous discussion and figures refer to representative simulations. How often do the RE algorithms tend away from the optimal action? An initial probability of attendance of .60 tends to favor attending, so here we consider the case where the optimal action is to attend. In Figure 10 the dotted lines show the data for one learner using the CS algorithm, the dashed grey lines show the data for the RE1 algorithm and the solid black lines for the RE2 algorithm. The solid line at .60 show the starting point for all the simulations. In all cases the probability of attendance for the other non-adaptive agents is below .60, so the optimal action for the learning agent is to always attend the bar. Figure 10 shows the data for different fixed probabilities of attendance of other agents. The lowest dotted line gives the probability of attendance at iteration 5000 for 100 different runs of the CS algorithm, with the outcomes ordered from lowest to highest. There are twenty-nine other agents with probability of attendance of .59. The next dotted line gives the same data when the probability of attendance for the other agents is .58, and so on though .55. The two highest lines, although not readily distinguished for the CS algorithm, show the data for .45 and .35. The lowest dashed grey line gives the results for the RE1 algorithm and the solid black line gives the results for the RE2 algorithm. When the probability of attendance is below .60 the process of learning led the agent to favor the action with the lower reward in the first 5000 iterations. For example, despite an initial condition that favored attending the RE1 algorithm moved away from the optimal action in about 35% of the simulations. The RE2 algorithm with "forgetting" performs better than the RE1 algorithm, but is still much less likely to take the correct action by time 5000 compared to the CS algorithm. For all three algorithms the lower the probability of attendance for the other agents, the higher the reward for attending the bar and the easier it is to discern the correct action.

4.2 Complex adaptive systems with multiple learners

When the number of adaptive agents increases, the external environment is not stationary and the actions of the adaptive agents interact over time. Again, the behavior of individual agents and of the system as a whole can differ dramatically depending on the initial conditions of the simulation. In some cases, the most notable feature of the majority of simulations is the rapid convergence of average attendance to .60, despite the relatively slow convergence of the individual adaptive agents. The learning rules considered here are much simpler specifications than the inductive learning approach that Arthur [1] utilized, but the generic behavior is remarkably similar. However, there are also cases where the interaction of the adaptive agents can lead to poor individual and system-wide performance.

Figures 11, 13 and 15 show 25000 iterations of the three algorithms with 15 learning agents and 15 agents who base their actions on independent realizations of a Bernoulli random variable with probability of attending of .75. (Figures 12, 14 and 16 are close-ups of figures 11, 13 and 15, respectively.) All of the other initial conditions are the same as the first example. The grey lines are the trajectories of the probability of attendance for the adaptive agents; the black line is the average probability of attendance for all agents including those with a fixed probability of attendance. The behavior of the individual algorithms is qualitatively similar to the case of a single adaptive agent, with the CS and RE2 algorithms converging towards a pure strategy and the RE1 algorithm rapidly approaching a relatively fixed probability of attending. The initial aggregate probability of attendance is .675. Mean attendance approaches 60% within 10 iterations for all algorithms and the variance of attendance declines over time, at least for the CS and RE2 algorithms. The greater variability of the CS algorithm leads to a greater variation of average attendance. The endogeneity of the environment also tends to slow down the convergence of the individual learners.

Table 1 summarizes the relationships between the average probability of attendance and the algorithm specifications when all agents are adaptive. Assuming approximately equal stepsizes or changes in probability across agents and across algorithms the largest change in

Table 1: Change in Average Probability of Attendance

learning rule	bar is uncrowded	bar is crowded
CS	$N \leq \mathcal{N}$ agents increase probability of attendance	$N > \mathcal{N}$ agents decrease probability of attendance
	$M-N > \mathcal{N}$ agents leave probability of attendance unchanged, if action of staying home has been taken several times	$M - N < \mathcal{N}$ leave probability of attendance unchanged, if action of staying home has been taken several times
net effect	increases	decreases
RE1	$N \leq \mathcal{N}$ agents increase probability of attendance	$N > \mathcal{N}$ agents leave probability of attendance unchanged
	$M - N > \mathcal{N}$ agents decrease probability of attendance	$M - N < \mathcal{N}$ decrease probability of attendance
net effect	indeterminate	decreases
RE2	$N \leq \mathcal{N}$ agents increase probability of attendance	$N \leq \mathcal{N}$ agents decrease probability of attendance
	$M - N > \mathcal{N}$ agents decrease probability of attendance	$M - N < \mathcal{N}$ decrease probability of attendance
net effect	indeterminate	decreases

the average probability is likely to occur when the bar is crowded bar and when CS agents attend an uncrowded bar. For the RE algorithms both actions are reinforced simultaneously when the bar is uncrowded. This suggests that there may be asymmetries between simulations which start with average attendance above and below .60%. The RE algorithms tend to be somewhat slower to converge (100–150 iterations) to an average probability of .60% when the initial probabilities of attendance are lower, but otherwise exhibit the same smooth aggregate behavior. The CS algorithm, on the other hand, performs poorly in an endogenous environment.

Figure 17 shows 5000 iterations of the CS algorithm with 15 learning agents and 15 agents who base their actions on independent realizations of a Bernoulli random variable with probability of attending of .45. The CS learners rapidly increase their probability of attendance until all 15 agents attend every period, the bar is crowded 96% of the time. Nonetheless it takes hundreds of iterations for the individual learners to begin to respond to the new environment. The scheme for averaging the rewards into weights is well adapted to a stationary environment but is slow to respond to new external conditions.

When all agents are adaptive the generic behavior of the RE algorithms is relatively rapid convergence of the aggregate probability of attendance despite the slow movement of individual agents' probabilities. The CS algorithm often rapidly overshoots aggregate attendance of .60%, for initial conditions above and below .60

Figures 20, 21 and 22 show the state of average attendance (large dots) and of 30 individual adaptive agents (small dots) after 5000 iterations. There were 25 separate simulations. The initial probability of attendance was again .60 for all agents. Consequently, these simulations are initialized at the symmetric mixed strategy Nash equilibrium. The mixed strategy equilibrium is fragile: the probability of attendance of individual adaptive agents rapidly moves away from .60. Note that the CS algorithm (figure 20) has recovered from its tendency to overshoot by iteration 5000 in these examples.

5 Conclusion

This paper examines the performance of simple learning rules in a complex adaptive system based on a coordination problem modeled on the *El Farol* problem. The key features of the *El Farol* problem are that it typically involves a medium number of agents and that agents' payoff functions have a discontinuous response to increased congestion. First we consider a single adaptive agent facing a stationary environment. We demonstrate that the simple learning rules proposed by Roth and Er'ev [17] and Er'ev and Roth [9] can be extremely sensitive to small changes in the initial conditions and that events early in a simulation can affect the performance of the rule over a relatively long time horizon. In contrast, a reinforcement learning rule based on standard practice in the computer science literature converges rapidly and robustly. The situation is reversed when multiple adaptive agents interact: the RE algorithms often converge rapidly to a stable average aggregate attendance despite the slow and erratic behavior of individual learners, while the CS based learners frequently over-attend in the early and intermediate terms. The symmetric mixed strategy equilibria is unstable: all three learning rules ultimately tend towards pure strategies or stabilize in the medium term at non-equilibrium probabilities of attendance. The brittleness of the algorithms in different contexts emphasize the importance of thorough and thoughtful examination of simulation-based results.

References

[1] W. B. Arthur, "Inductive reasoning and bounded rationality: The *El Farol* problem", *American Economic Review: American Economic Association Papers and Proceedings 1994*, Vol. 84, Pp. 406-411, May 1994.

[2] A. M. Bell, W. A. Sethares and J. A. Bucklew, "Coordination failure as a source of congestion in information networks," mimeo, NASA Ames Research Center, http://ic.arc.nasa.gov/ic/people/bell/bell.html, May 1999.

[3] T. Borgers and R. Sarin, "Learning through reinforcement and replicator dynamics," mimeo, University College, London, 1995.

[4] T. Borgers and R. Sarin, "Naive reinforcement learning with endogenous aspirations," mimeo, University College, London, 1996.

[5] R. Bush and F. Mosteller, *Stochastic models for learning*, New York: Wiley, 1955.

[6] J. L. Casti, "Seeing the light at *El Farol*," *Complexity*, Vol. 1, No. 5, Pp. 7-10, 1996.

[7] D. Challet and Y.-C. Zhang, "On the minority game: Analytical and numerical studies," *Physica A*, Vol. 256, pp. 514-532, 1997.

[8] B. Edmonds, "Gossip, sexual recombination and the *El Farol* bar: Modeling the emergence of heterogeneity," *Proceedings of the 1998 Conference on Computation in Economics, Finance, and Engineering*, Cambridge, England, June 1998.

[9] I. Er'ev and A. E. Roth, "Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria," *American Economic Review*, Vol. 88, No. 4, pp. 848-881, September 1998.

[10] D. Fudenberg and D. K. Levine, *The Theory of Learning in Games*, The MIT Press: Cambridge, Massachusetts, 1998.

[11] N. S. Glance and B. A. Huberman, "The dynamics of social dilemmas," *Scientific American*, pp. 76-83, March 1994.

[12] A. Greenwald, Bud Mishra, and Rohit Parikh, "The Santa fe bar problem revisited: Theoretical and practical implications," mimeo, New York University, New York, NY, 1998.

[13] B. A. Huberman and R. M. Lukose, "Social dilemmas and Internet congestion," *Science*, Vol. 277, pp. 535-537, July 25, 1997.

[14] N. F. Johnson, S. Jarvis, R. Jonson, P. Cheung, Y. R. Kwong and P. M. Hui, "Volatility and agent adaptability in a self-organizing market," *Physica A* Vol. 256, No. 230, 1998.

[15] D. Monderer and L. Shapley, "Potential games," *Games and Economic Behavior*, No. 14, pp.124-143, 1996.

[16] R. Rosenthal, "A class of games possessing pure strategy Nash equilibria," *International Journal of Game Theory*, No. 2, pp. 65-67, 1973.

[17] A. E. Roth and I. Er'ev, "Learning in extensive form games: Experimental data and simple dynamic models in the intermediate term," *Games and Economic Behavior*, No. 8, pp. 164-212, 1995.

[18] R. Savit, R. Manuca, Y. Li and R. Riolo, "The dynamics of minority competition," in *Proceedings of the Second International Conference on Complex Systems*, Nashua, New Hampshire, October 1998.

[19] W. A. Sethares and A. M. Bell, "An adaptive solution to the *El Farol* problem," Proc. of the Thirty-Sixth Annual Allerton Conference on Communication, Control, and Computing, Allerton IL, Sept. 1998.

[20] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, (Cambridge, MA: The MIT Press), 1998.

[21] D. Wolpert and K. Tumer, "An introduction to collective intelligence," forthcoming in *The Handbook of Agent Technology*, Jeffrey M. Bradshaw, editor, AAAI Press/MIT Press, 1999.

[22] D. Wolpert, K. Wheeler and K. Tumer, "Collective intelligence for control of distributed dynamical systems," mimeo, NASA Ames Research Center, http://ic.arc.nasa.gov/ic/people/kagan/coin_pubs.html, 1998.

[23] E. Zambrano, "Rationalizable boundedly rational behavior", *Sania Fe Institute Working Paper*, No. 97-06-060, 1997.

Figure 1: Probability of attendance for one adaptive agent using the CS (bottom line), RE1 (top line) and RE2 (middle line) algorithms. The fixed probability of attendance for non-adaptive agents is .61. The optimal action is to stay home every period.

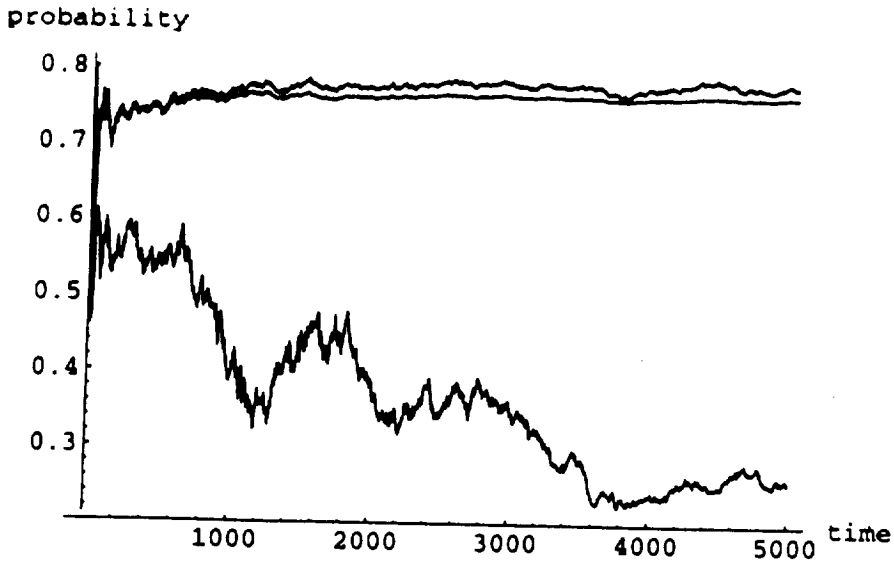

Figure 2: Probability of attendance for one adaptive agent using the CS, RE1 and RE2 algorithms. First 5000 iterations of figure 1.

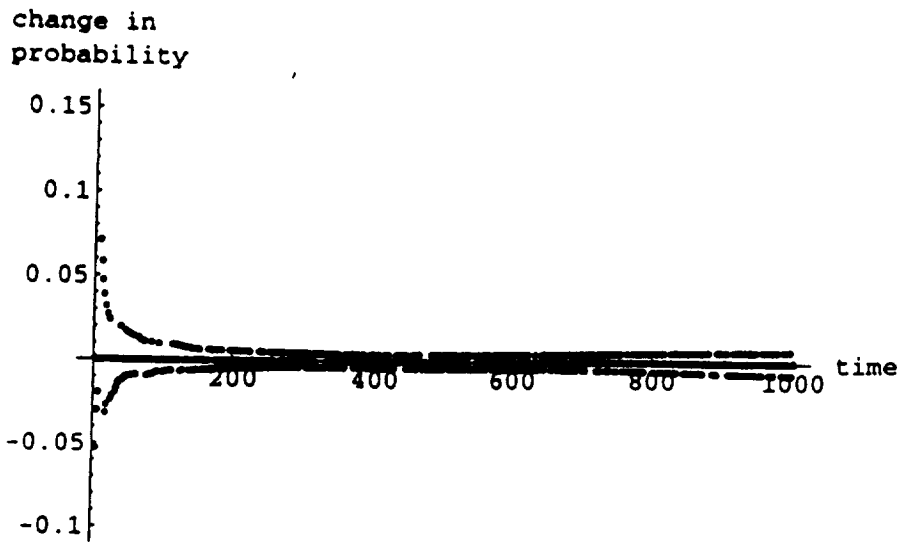

Figure 3: Change in probability of attending for the CS algorithm. First 1000 iterations of figure 1.

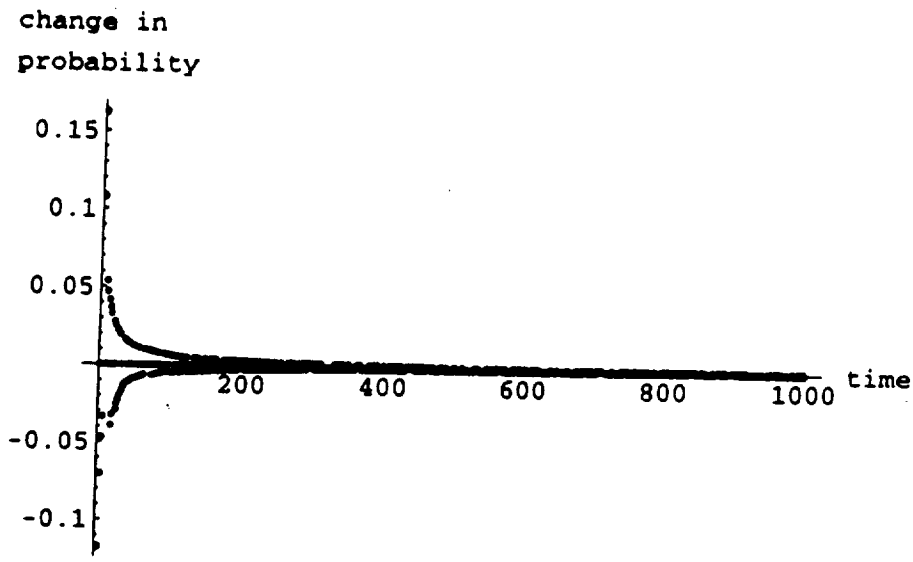

Figure 4: Change in probability of attending for the RE1 algorithm. First 1000 iterations of figure 1.

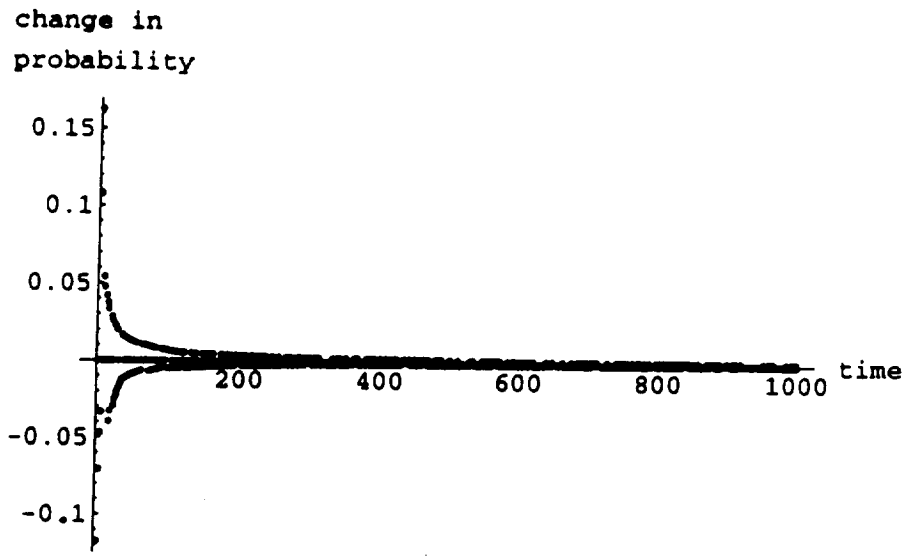

Figure 5: Change in probability of attending for the RE2 algorithm. First 1000 iterations of figure 1.

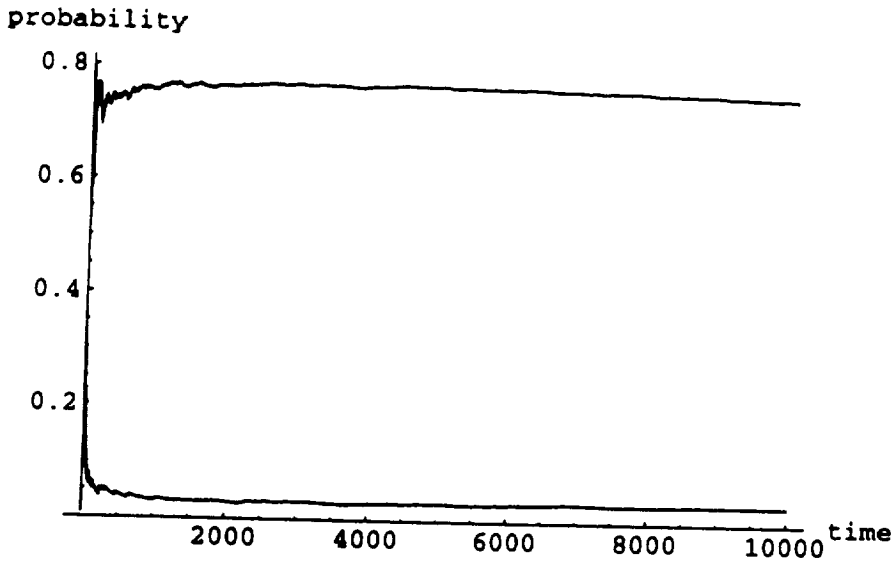

Figure 6: Probability of attendance for an adaptive agent using the RE1 algorithm with initial weights of $\{.8, 1.0411\}$ (upper solid line) and initial weights of $\{.8, 1.04105\}$ (lower dashed line). The fixed probability of attendance for non-adaptive agents is .61. The optimal action is to stay home every period.

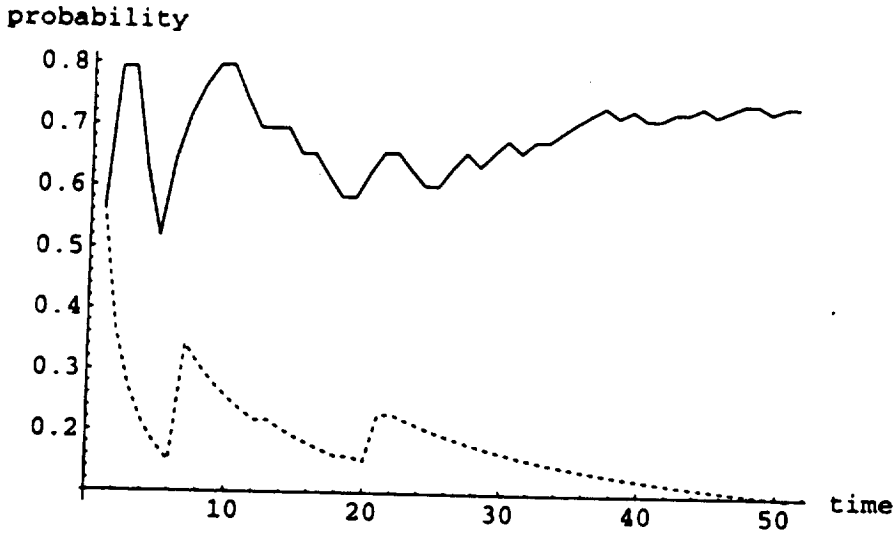

Figure 7: Probability of attendance for an adaptive agent using the RE1 algorithm with initial weights of $\{.8, 1.0411\}$ (upper solid line) and initial weights of $\{.8, 1.04105\}$ (lower dashed line). First 50 iterations of figure 6.

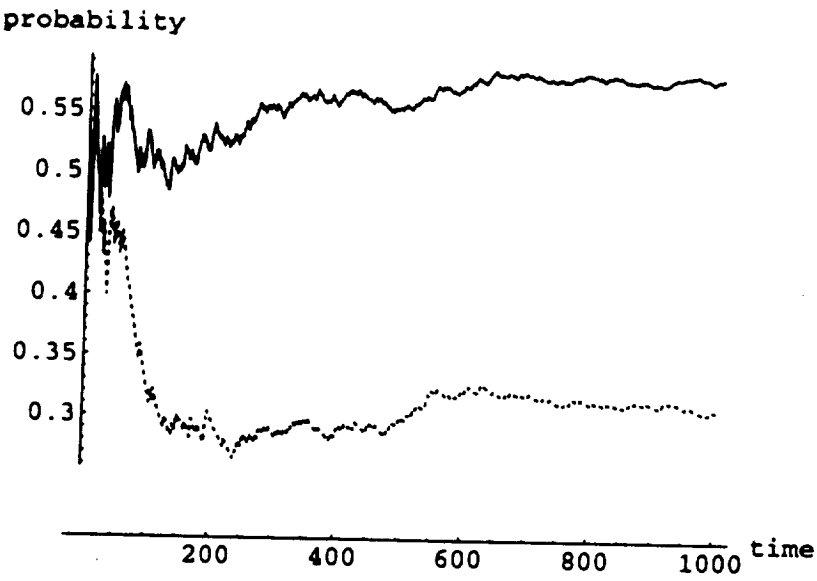

Figure 8: Probability of attendance for an adaptive agent using the RE1 algorithm with initial weights of $\{8, 9\}$ (upper solid line) and initial weights of $\{8, 8.75\}$ (lower dashed line).

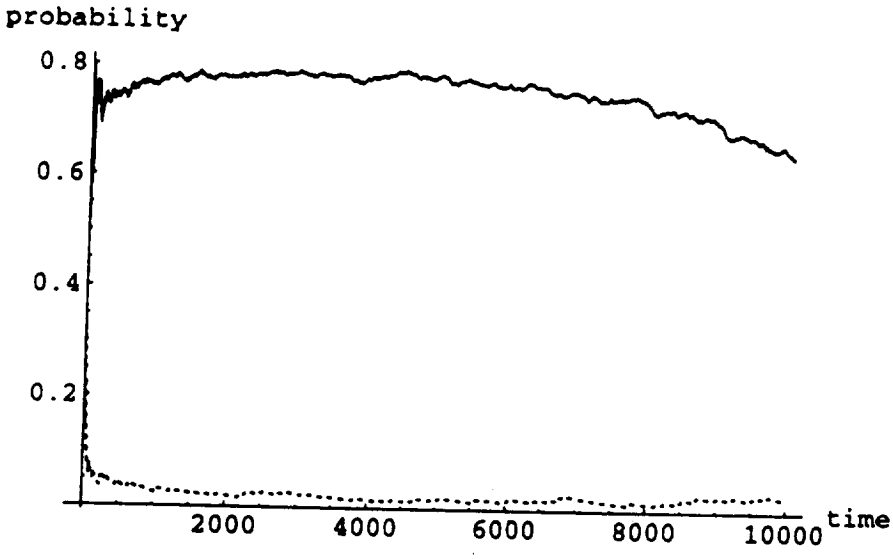

Figure 9: Probability of attendance for an adaptive agent using the RE2 algorithm with initial weights of $\{.8, 1.05\}$ (upper solid line) and initial weights of $\{.8, 1.04\}$ (lower dashed line). The fixed probability of attendance for non-adaptive agents is .61. The optimal action is to stay home every period.

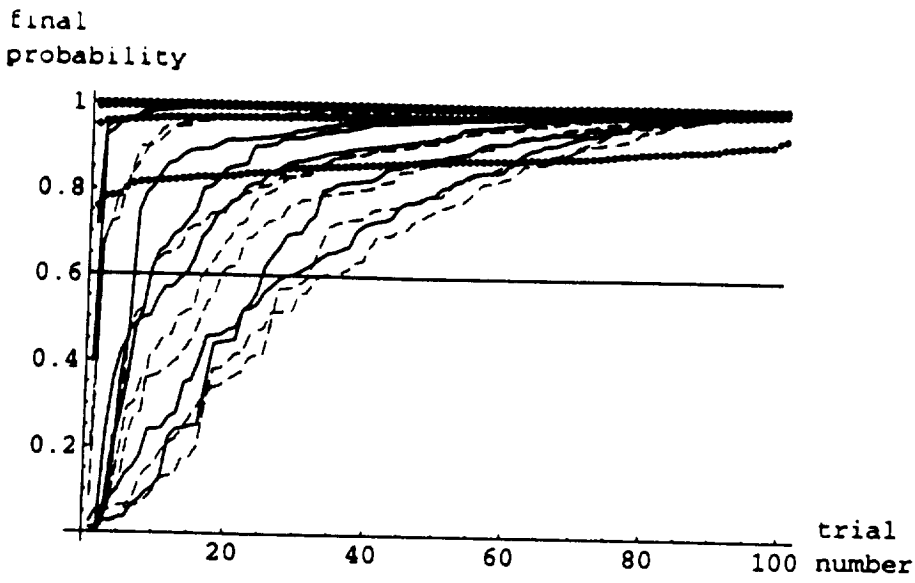

Figure 10: Probability of attendance for one adaptive agent using the CS (dotted line), RE1(dashed line) and RE2 (solid line) algorithms with 29 non-adaptive agents. Each line represents 100 simulations with differing fixed probabilities of attendance for the non-adaptive agents, ranging from .59 to .35.

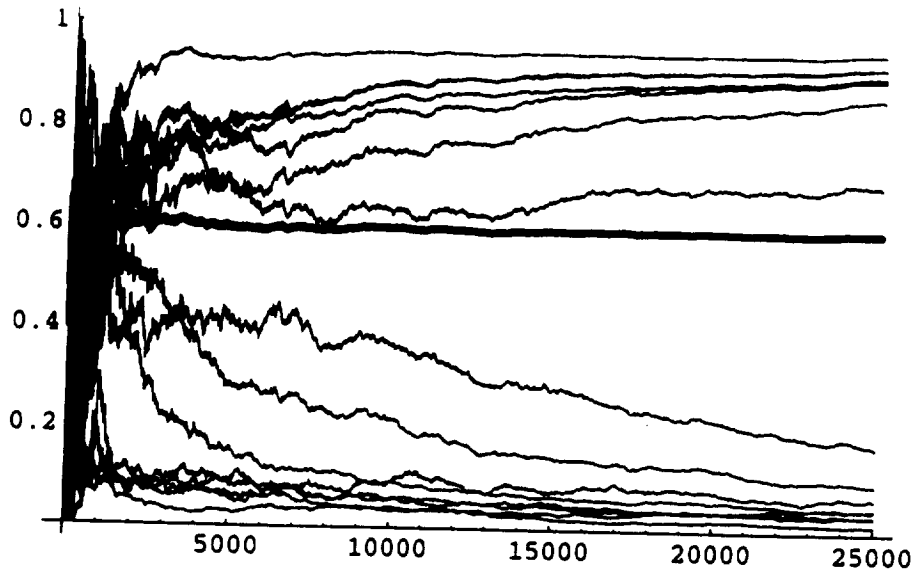

Figure 11: Probability of attendance for 15 adaptive agents (grey lines) using the CS algorithm with 15 non-adaptive agents with probability of attendance of .75. Average probability of attendance for all agents is shown by the solid black line.

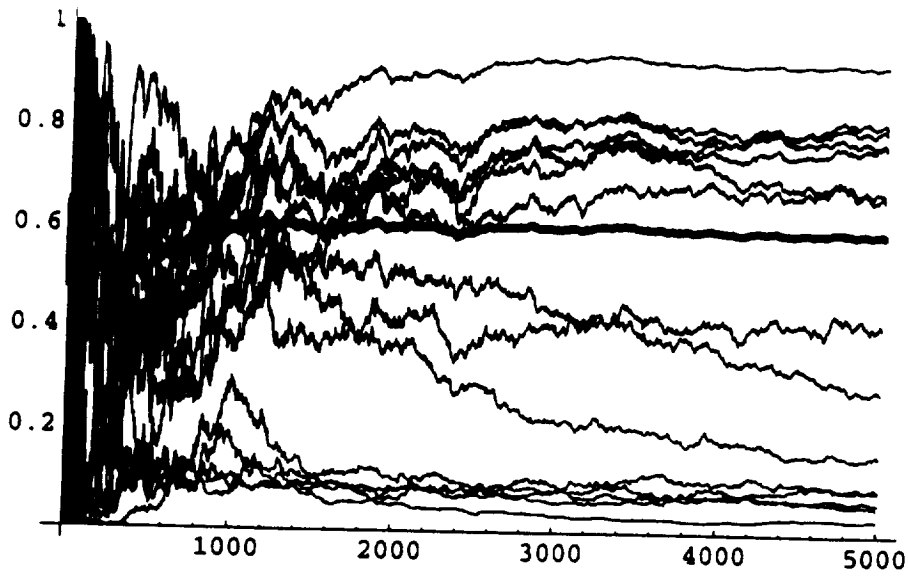

Figure 12: Probability of attendance for 15 adaptive agents (grey lines) using the CS algorithm with 15 non-adaptive agents with probability of attendance of .75. First 5000 iterations of figure 11.

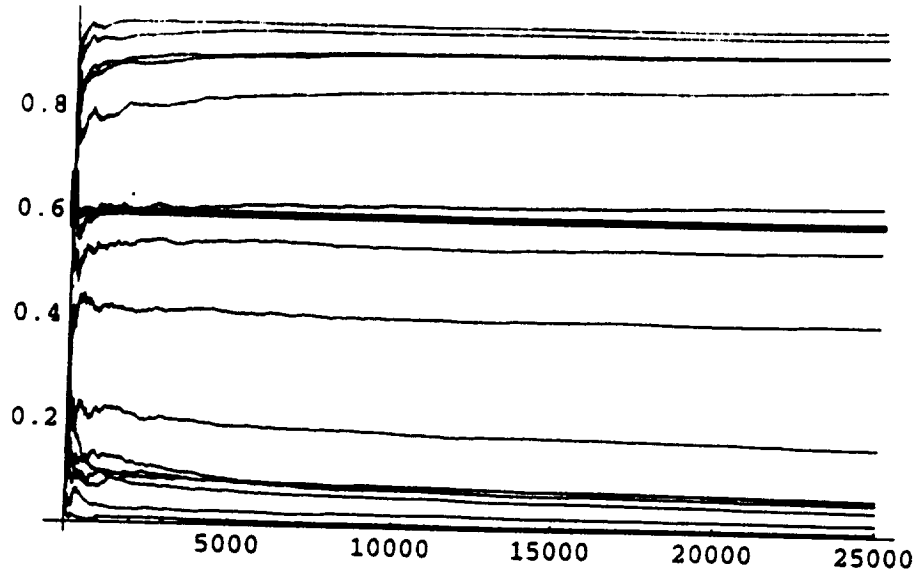

Figure 13: Probability of attendance for 15 adaptive agents (grey lines) using the RE1 algorithm with 15 non-adaptive agents with probability of attendance of .75. Average probability of attendance for all agents is shown by the solid black line.

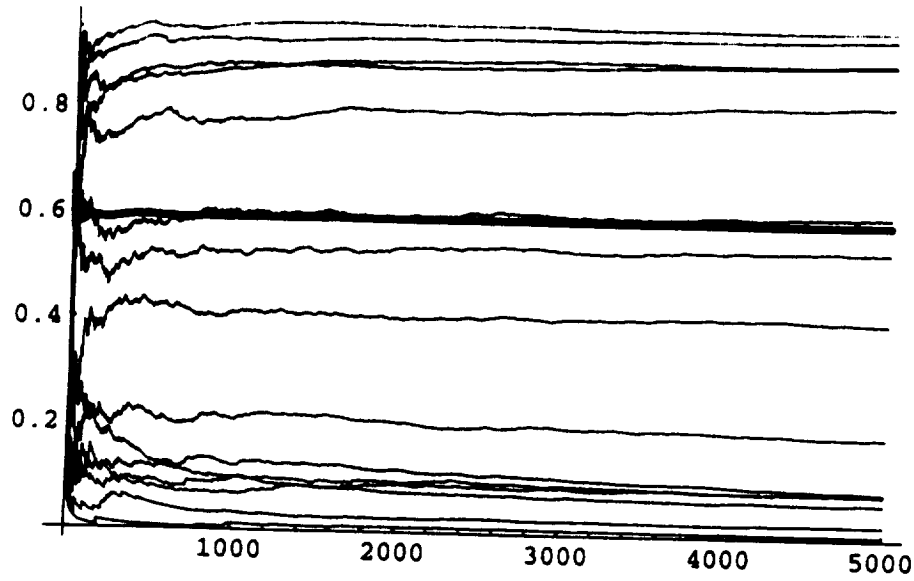

Figure 14: Probability of attendance for 15 adaptive agents (grey lines) using the RE1 algorithm with 15 non-adaptive agents with probability of attendance of .75. First 5000 iterations of figure 13.

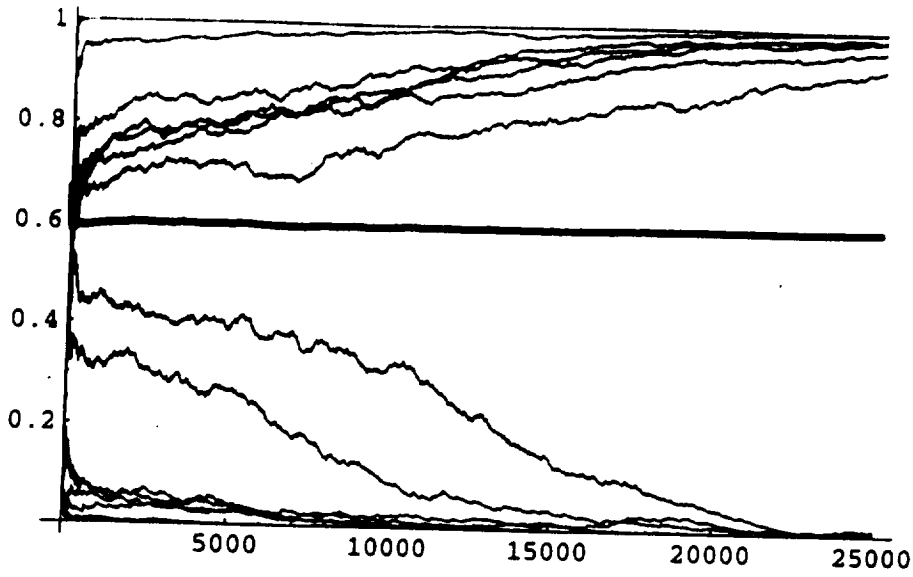

Figure 15: Probability of attendance for 15 adaptive agents (grey lines) using the RE2 algorithm with 15 non-adaptive agents with probability of attendance of .75. Average probability of attendance for all agents is shown by the solid black line.

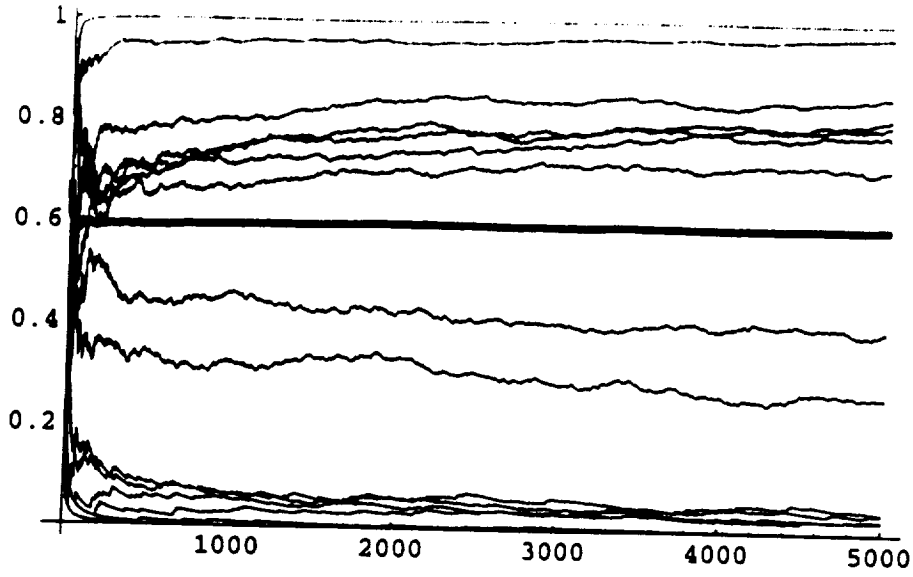

Figure 16: Probability of attendance for 15 adaptive agents (grey lines) using the RE1 algorithm with 15 non-adaptive agents with probability of attendance of .75. First 5000 iterations of figure 15.

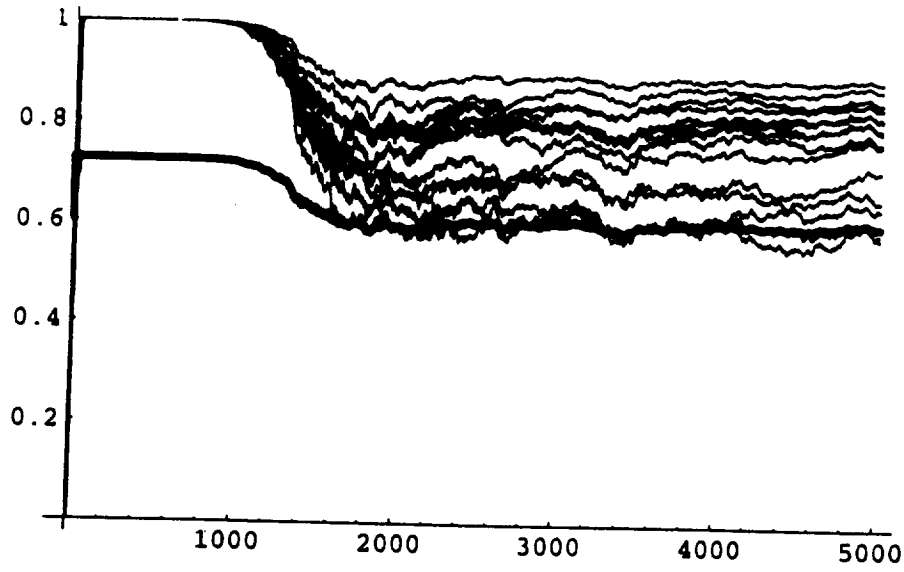

Figure 17: Probability of attendance for 15 adaptive agents (grey lines) using the CS algorithm with 15 non-adaptive agents with probability of attendance of .45. Average probability of attendance for all agents is shown by the solid black line.

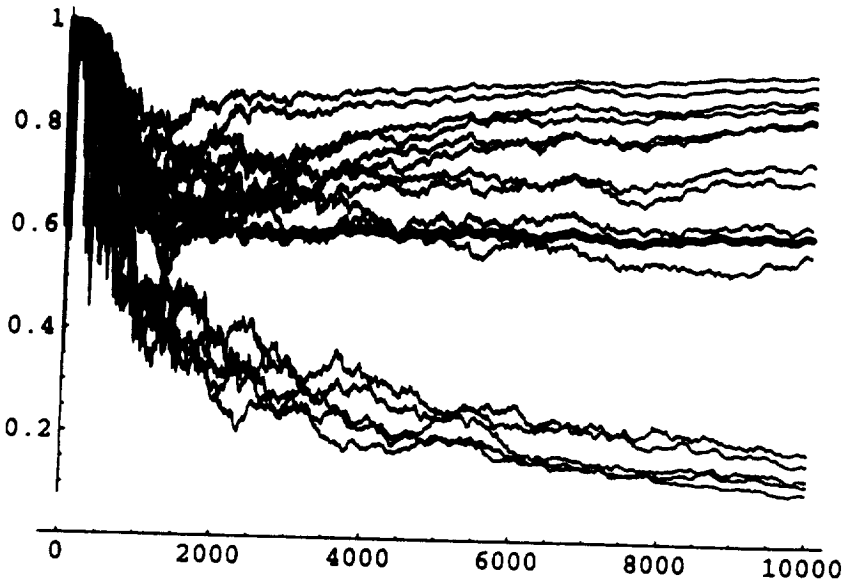

Figure 18: Probability of attendance for 30 adaptive agents (grey lines) using the CS algorithm with initial probability of attendance of .60. Average probability of attendance for all agents is shown by the solid black line.

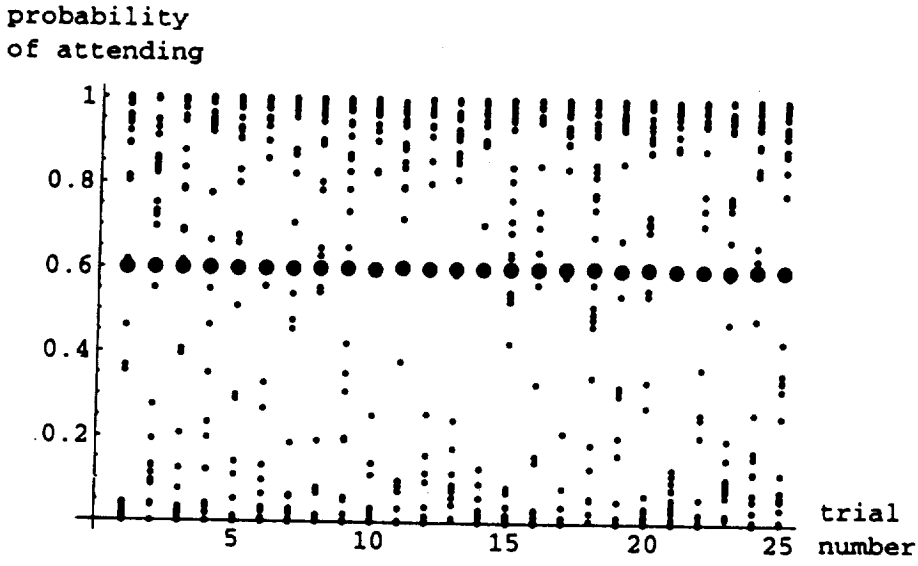

Figure 22: Probability of attendance for 30 adaptive agents (small dots) using the RE2 algorithm. Average probability of attendance for all agents is shown by the large dots. Each vertical column of dots presents the data from one simulation.