Dynamics of strategy distribution in iterated games

Stéphane Airiau and **Sabyasachi Saha** and **Sandip Sen**
Department of Mathematical & Computer Sciences
University of Tulsa
Tulsa, Oklahoma, USA
{stephane, saby, sandip}@utulsa.edu

Abstract

Evolutionary tournaments have been used as a tool for comparing game-playing strategies. For instance, in the late 1970's, Axelrod organized tournaments to compare strategies for playing the iterated prisoner's dilemma (PD) game. While these tournaments and later research have provided us with a better understanding of successful strategies for iterated PD, our understanding is less clear about strategies for playing iterated versions of arbitrary single-stage games. While solution concepts like Nash equilibria has been proposed for general-sum games, learning strategies like fictitious play may be preferred for playing against sub-rational players. In this paper, we discuss the relative performance of both learning and non-learning strategies in different population distributions including those that are likely in real-life. The testbed used to evaluate the strategies includes all possible structurally distinct 2×2 conflicted games with ordinal payoffs. Plugging head-to-head performance data into an analytical finite-population evolution model allows us to evaluate the evolutionary dynamics of different initial strategy distributions. Two key observations are that (a) the popular Nash strategy is ineffective in most tournament settings, (b) simple strategies like best response benefit from the presence of learning strategies and we often observe convergence to a mixture of strategies rather than to a single dominant strategy. We explain such mixed convergence using head-to-head performance results.

Introduction

Learning and reasoning in single or multistage games have been an active area of research in multiagent systems (Bowling & Veloso 2001; Claus & Boutilier 1998; Littman 1994). In particular, iterative versions of single-stage bimatrix games have been used to evaluate learning strategies by multiagent researchers. Particular games like the Prisoner's Dilemma (PD) have received widespread attention both in game theory and in multiagent systems. Solution concepts like Nash Equilibria (NE) has been propounded as desired goals for rational play though there exists several criticism of this view. Though it follows from its definition that an opponent of a Nash player cannot do better than playing its component of NE, playing a Nash strategy is not necessarily the best option against a non-Nash player. A learning strategy that tries to predict the move of the opponent and optimally responds to that may be a better option against sub-rational players.

We are interested in comparing learning and non-learning strategies on a standardized set of games against a possible collection of opponents. The testbed we adopted is a set of all structurally distinct conflicting 2x2 games with ordinal payoffs (Brams 1994). We assume that players have complete information, i.e. each player is aware of both its own and its opponent's payoff matrix. We described the testbed in our earlier paper where we focused on evaluating strategies in a round robin tournament (Airiau & Sen 2003). The results presented then clearly showed that learning strategies are outperforming non-learning strategy. Each strategy was represented by one player, assuming the strategy distribution over a population is static and uniform, which was unrealistic. To address this problem, we study the results of evolutionary tournaments in a potentially large, but finite, population. We consider a fixed population of agents. In each generation, the agents are playing a round robin tournament using a fixed strategy. Based on the results of the tournament, each agent can change strategy for the next generation. This evolutionary setting reflects the desire of a rational agent to adopt a better performing strategy. We develop a finite-population analytical model for capturing this evolutionary process and study the population dynamics for different initial agent strategy distributions.

In this paper, we are first going to present the representative strategies chosen. Then, we present the evolutionary tournament in which round robin matches are played to evaluate head to head performance and after, a selection mechanism is used to generate strategy distribution of the next generation. In the next section, we present an analytical model of our selection mechanism. Finally, we discuss the outcome of evolutionary tournaments with different initial strategy distributions, making different assumptions about the sophistication of the agents in the population.

Strategies

We chose the strategies used in our tournament from well-known learning and non-learning strategies (and one that was the winner in a local competition between students):

Random: The action played is chosen from an uniform distribution over its action space. The use of this strategy can also model a collection of other strategies represented in the population.

MaxiMin(M): The action chosen is the one that produces maximum lower bound payoff.

Nash(N): One of the Nash equilibrium strategies (Nash 1951) is played. A strategy combination $(\pi_1, \ldots, \pi_n)$ is in Nash Equilibria (NE) if $\forall i$, $r_i(\pi_1, \ldots, \pi_i, \ldots, \pi_n) \geq r_i(\pi_1, \ldots, \pi'_i, \ldots, \pi_n)$, where $r_k(\pi_1, \ldots, \pi_n)$ is the payoff of player k and π'_i is any other valid strategy for i. This means at NE, no player has incentive to unilaterally deviate from its current strategy. For non-communicating rational players a strategy combination at NE is stable. To compute the different Nash equilibria for the games, we used Gambit[1]. Out of the 57 games used in the testbed, 6 games have multiple Nash equilibria. Since it is unclear how non-communicating Nash players will choose from multiple equilibria, we randomly selected the Nash equilibrium played.

Tit for tat (TFT): This strategy is famous in the context of the prisoner's dilemma and the tournament ran by Axelrod (in this strategy, the player will play cooperate if and only if the opponent played cooperate in the previous iteration, hence the name "tit for tat"). In the context of our tournament, a player using the tit for tat strategy will play the action that the opponent played during the previous iteration. This strategy is purely reactive and takes into account only the previous decision of the opponent.

Best Response to previous action (BR): A (BR) player can be viewed as a sophisticated TFT player: instead of playing the last action i of the opponent, the player responds with the best response to i. In other words, the player playing the best response strategy assumes that its opponent is playing a pure strategy and answers optimally to it. BR is also purely reactive and models the opponent as a player either using a pure strategy or one with a strong sense of inertia, i.e. aversion to change.

Fictitious Play (FP): This is the basic learning approach well-known in game theory literature (Fudenberg & Levine 1998). The player keeps a frequency count of its opponent's decisions from a history of past moves and computes the mixed strategy being played by its opponent. It then chooses its best response to that mixed strategy, with the goal of maximizing expected payoff. This player models its opponent's behavior and tries to respond in an optimal way. If the opponent is playing a fixed pure or mixed strategy, FP will be able to respond optimally.

Best response to Fictitious play (BRFP): This strategy assumes that the population is composed of many learning agents using the FP strategy. The player models its opponent as a FP player: knowing its own history of actions, it can determine what an agent using FP would do, and it computes the best response to this action. We incorporated this strategy assuming that given that FP is a reasonable learning strategy to play, a player can choose to adopt a strategy to respond optimally to FP.

[1] http://www.hss.caltech.edu/gambit

Saby: The last strategy that we have used was the one that won a local tournament between students in a multiagent systems course. This learning strategy assumes that the opponent is likely to respond to my moves and tries to model the probability distribution of the opponent's moves given my last move. This is akin to a 2-level player compared to a 1-level player in our prior work (Mundhe & Sen 2000). For its own action i, in the last time period, the agent first calculates the conditional probability of action k of the opponent to be proportional to the average utility the opponent received for choosing action k the last t times it played k when this player played i in the previous time step. These numbers are normalized to obtain the conditional probabilities the opponent are expected to use in choosing action in the next iteration. The agent then plays a best response to that probability distribution.

We believe that probably not all of these strategies would be used in an open environment. It seems reasonable to assume that simple strategies such as R, TFT, BR and M would be used. Because of the popularity of the concept of the Nash equilibrium and as the basic learning approach, Nash and FP are also likely to be used. We consider Saby as strategy that is used by a minority of players. We did not consider pure strategy players, i.e., players who always chose a specific action, as the semantics of any action varies considerably over the different games.

In our study, we are interested in two criteria for comparing the strategies: the complexity of the strategy and whether learning is used.

Simple Vs Complex strategies: Random (R), Tif For Tat (TFT), Best Response (BR) and MaxiMin (M) are considered to be simple strategies. The random strategy can be interpreted as the ensemble of behavior of a collection of different lesser known strategies as well as behavior exhibited by inconsistent players. On the other hand, We hypothesize that playing Nash equilibrium (N) is a complex strategy since computation of a Nash is NP complete. Also, fictitious play (FP), Best Response to FP and Saby are considered to be complex strategy

Learning Vs Non-learning strategies: Random, Nash and MaxiMin are static strategies which do not respond to the opponent. TFT and BR are simple, purely reactive strategies, that can be considered as a primitive learning strategies: an agent using TFT mimics the last action of the opponent. Instead of mimicking the last action, an agent using BR plays the best response to this action. The remaining strategies are learning strategies. The strategy FP is the basic learning approach. If we assume that many agents are using this basic learning approach, it is possible to use a strategy which plays optimally against FP, hence the use of BRFP. We introduced Saby strategy which also uses learning.

Tournament Structure

In this section, we describe the underlying tournament structures given the set of matrices and a selection of strategies.

Round Robin Play

Each player has complete information about the game, including the payoff matrices of its own as well as that of its opponent. The players are not allowed any other means of communication apart from expressing their action at each iteration.

All agents in the population play with all other agents. This round robin form of play allows us to obtain a head-to-head performance between any two strategies and also to compute relative performance of any agent given an arbitrary strategy mix in the population. In round robin play, each player plays with each of the other players and itself over all 57 matrices of the testbed. The exhaustive set of 57 possible type of matrices represent all the distinct conflicting situation with ordinal payoffs. 51 of these games have a unique Nash equilibrium (9 of these games have a mixed strategy equilibrium and 42 have pure strategy equilibrium), the remaining 6 have multiple equilibria (two pure and a mixed strategy). Of the 42 games that have a unique pure strategy Nash equlibrium, in 4 games the Nash equilibrium is not pareto-optimal. To eliminate the bias of the construction of the matrices (playing as a column player is preferable), each player plays every other player both as a column and as a row player for each of the matrices. To collect meaningful results, each game is iterated 100 times. Because the action space is small, we assumed that 100 iterations are reasonable for players that use a learning approach to adapt their strategies. To evaluate the stable performance of the players, we accumulate the payoffs of the players only over the last 50 iterations of the game. The score of one player is the cumulative score obtained over all the games played against all other players.

Evolutionary Tournament

The evolutionary tournament is run over a fixed population of agents. During each generation, agents in the population engage in round robin play and do not change their strategy. We assume that players have no prior knowledge of the strategy used by its opponent during a game. However, at the end of round robin play, the players can observe the strategy used by all the other players and their scores. Based on this information, the agents can then decide to change their strategy. We assume that they have knowledge of and can execute all strategies. We have used a modification of the tournament selection algorithm (Deb & Goldberg 1991) to determine new strategies to be used by the agents in the next generation. As described in Algorithm 1, each agent picks two agents with a probability proportionate to their score (which has a flavor of fitness proportionate selection), and then it decides to adopt the strategy of the better of the two (tournament selection). This variant promotes strategies which are doing well in the population and corresponds to realistic scenarios where it is more likely that relatively successful agents will be noticed and their behavior imitated by others in the population. Note that this particular form of selection produces an even stronger selection bias for higher performing individuals than produced by fitness proportionate (because it does not allow head to head comparisons, small absolute differences are not recognized) or tournament (because parents are picked randomly rather than being biased by performance) selection alone.

Algorithm 1 Tournament Selection Algorithm.

$strat(i)$ denotes the strategy of player i
$score(i)$ denotes the cumulative results obtained by player i during one instance of the tournament
for N iterations **do**
  for every player k **do**
    $$Prob(pick\ k) = \frac{score(k)}{\sum_i score(i)}$$
  for every player k **do**
    pick randomly ρ_0 according to Prob
    pick randomly ρ_1 according to Prob
    $newstrat(k) \leftarrow strat(argmax_{i\in\{0,1\}}(score(\rho_i)))$
  for every player k **do**
    $strat(k) = newstrat(k)$

A strategy which performed well in a generation is likely to increase in proportion at the expense of below-average strategies,

Analytical model of the selection mechanism

We have developed an analytical model for a finite population which outputs the strategy distribution of the next generation, given the current strategy distribution, the head to head results between the strategies and the number of agents. The model saves us the cost of running Algorithm 1 to obtain the dynamics of the system.

Let I denote the number of strategies in the domain. In our paper, we have $I = 8$. Let N_i be the number agents of type i, and we have $N = \sum_{i=1}^{I} N_i$, is the total number of agents. Let $U_{i,j}$ denotes the average payoff received by agent of type j playing against agent of type i. This average is computed over 10 round robin tournaments over a population containing one agent per strategy (see Section). Given this result, we can compute the average payoff S_i received by an agent of type i playing against all other agents including itself:

$$S_i = \sum_{1\leq j\leq I} \frac{N_j}{N} * U_{i,j}.$$

Let $P(l,i)$ denote the probability that agent l will choose to evolve to an agent of type i in the next generation and $N^1(i)$ denote the number of agents of type i in the next generation. Then, since the agent's decision is not dependent of other agent's decision, we have

$$N^1(i) = \sum_{l=1}^{N} P(l,i) = N \times P(l,i).$$

$P(l,i)$, the probability that agent l will choose strategy i, is the sum of two probabilities in the tournament selection. The first is the probability that l picks two agents of the same type i in the tournament selection, in which case the i^{th} strategy will be chosen with certainty. The second

Rank	Player	average score per game
1	Saby	2.99
2	BRFP	2.98
3	FP	2.96
4	BR	2.94
5	Nash	2.94
6	MaxMin	2.81
7	TFT	2.75
8	R	2.44

Table 1: Strategy ranking from round robin play, one player per strategy. Average score is the average over all games played with other players.

is the probability of choosing one agent with strategy i and the other agent with strategy j such that $j \neq i$ and $S_i > S_j$. If $S_i = S_j$ then i is chosen with probability 0.5. Hence we have:

$$P(l,i) = C(i,i) \times 1 + \sum_{j=1,\neq i}^{I} C(i,j) \times \{P(S_i > S_j) + P(S_i = S_j) \times 0.5\}$$

where, $C(i,j)$ is the probability that an agent with strategy i and an agent with strategy j are chosen.

$$C(i,j) = \begin{cases} \frac{N_i f_i \times (N_i - 1) f_i}{N(N-1)\overline{f}^2} & \text{if } j = i \\ \frac{2 \times N_i f_i \times N_j f_j}{N(N-1)\overline{f}^2} & \text{otherwise,} \end{cases}$$

where f_i and $\bar{f}$ are average score by the agents with strategy i and average score of the all agents.

We define $P(S_i > S_j)$ to be the probability that an agent with strategy i will have better score compared to an agent with strategy j. We approximate it as the proportion of times agents with strategy i have done better compared to the agents with strategy j. Using the values S_i computed, we can find the values of the probabilities $P(S_i = S_j)$ and $P(S_i > S_j)$.

Among other things the analytical model provides us the following capabilities:

- It enables us to greatly reduce the cost of calculating the outcome of the evolutionary process by substituting actual tournaments with simple calculations.
- It enables us to compute outcomes for arbitrary initial configurations and arbitrary strategies given head-to-head results between these strategies.

Results

Head to Head results among the strategies

The results of round robin play in a population with one player per strategy provides us with an unbiased relative performance of the strategies. The rankings of the different strategies together with their payoffs, averaged over all interactions, are presented in Table 1. Learning strategies are performing better than non learning strategies in this setting. Analysis of comparable results (without the MinMax player) can be found in (Airiau & Sen 2003). Though these results are interesting, they do not provide significant insight about the dynamics of a population when agents are allowed to change strategies, leading to a non-uniform distribution of agents.

To better understand the dynamics of the evolutionary tournaments, we present in Table 2 the head to head results. Each entry is the average score of the row player when it played against the column player. An entry in the diagonal is the result of self play, which is critical in a population where many agents use the same strategy. The relative head-to-head performance of the strategies are recorded in Table 3. Each entry is the difference between the row player and the column player: if the entry is positive, the row player wins the head to head confrontation.

Evolutionary Tournament results

We have studied the evolution of various populations with different strategy distribution. We have focused on two main population types based on the constituent strategies: the first contains agents with only simple strategies, the second contains all of the representative strategies.

We have also studied the effect of adding more sophisticated agents (N and FP for the simple population, and BRFP and Saby for the other population) to these populations. We believe that in the real-world, at least at the outset, more sophisticated strategies are likely to be used by only a minority of players.

All the figures have been generated using the model presented in Section . We have compared the model with actual runs of the evolutionary tournament and the corresponding population dynamics match closely. The only, relatively infrequent, mismatches resulted from sampling of strategies with extremely low selection probabilities.

Population of simple agents We first consider a population of simple agents that can use either of the following strategy: Random, TFT, BR, and M. Though one can argue, we consider TFT to be a rudimentary learning scheme since the strategy mimics the behavior of the opponent. BR can be seen as the next step of the logical progression and a slightly more sophisticated 'learning' method since it predicts that the opponent will repeat its last move and hence plays the best response to that. Interestingly, the population converges to the use of this strategy, as shown in Figure 1, even when only a single agent is playing BR and each of the other strategies, i.e., TFT, Random and MaxMin is used by 1000 agents. The proportions of Random and TFT decrease rapidly due to the presence of MaxMin (referring back to 3, we find that MM gains a lot at the expense of R and TFT). BR can exploit both R and MM agents but is exploited by the TFT agents. Only when the TFT agents are eliminated from the population that the proportion of BR starts to rise. When we experimented with an initial uniform distribution of these strategies, the convergence to BR is faster.

The above scenario conforms to the following general

	R	TFT	N	BR	FP	BRFP	MM	Saby
R	2.539766	2.5520468	2.383041	2.45731	2.4391813	2.3766081	2.3169591	2.4666667
TFT	2.5140352	2.588304	2.916959	2.8543859	2.7251463	2.888304	2.5730994	2.9538012
N	2.937427	2.9321637	2.939766	2.9444447	2.9356725	2.9614034	2.9076023	2.962573
BR	2.883041	2.732164	2.9397662	2.9157894	3.011111	2.940351	3.0695906	3.031579
FP	2.9660819	2.9883041	2.908772	2.974854	2.9532166	2.9099417	3.0391812	2.9315789
BRFP	2.8818712	2.849123	2.9245615	3.1163745	3.1777778	2.9239767	2.852047	3.097076
MM	2.9894738	3.1415205	2.7146199	2.71462	2.7479534	2.7678363	2.631579	2.7608187
Saby	2.9128656	3.0204678	2.930994	2.9988303	3.0649123	2.94269	3.0105262	3.0005846

Table 2: Head to head results.

TFT	-0.03801155						
N	0.55438614	0.015204668					
BR	0.42573094	-0.12222195	-0.004678488				
FP	0.52690053	0.26315784	-0.02690053	-0.03625703			
BRFP	0.5052631	-0.039180994	-0.03684187	0.17602348	0.2678361		
MM	0.6725147	0.5684211	-0.19298244	-0.35497046	-0.29122782	-0.084210634	
Saby	0.44619894	0.0666666	-0.031579018	-0.0327487	0.13333344	-0.15438604	0.24970746
	R	TFT	N	BR	FP	BRFP	MM

Table 3: Relative performance: difference between the head to head score of the row player and the column player.

trend observed in a number of other scenarios: the initial dominance of one strategy, ID, eliminates the set of strategies it exploits the most, $E(ID)$, allowing the emergence of some ultimately dominant strategy, UD, such that the following conditions hold:

- $\exists S | (S \in E(ID)) \wedge (UD \in E(S))$, i.e., there exists some strategy that is dominated by ID and in turn dominates UD.
- $ID \in E(UD)$, i.e., ID is exploited by UD,

As we typically do not have strategies that exploit all strategies or are not exploited by any of the other strategies, we do not find monotonic growth of one strategy that takes over the population. Rather, we have the more complex scenario of an early dominant strategy eliminating the obstacles to the ultimate dominance of another strategy. So, the early winner, in effect, unknowingly creates an environment congenial for its own failure and ultimately extinction. In some later cases we see not an extinction but a see-saw “battle” of survival between UD and ID. Such a situation arises when one of these strategies dominates the other only when they are in the minority in the population! In these cases, we observe a cyclical behavior in the system, with the UD and the ID strategies taking the upper hand in successive generations.

In the second set of experiments, we introduce 10 Nash agents in a population containing 1000 agents each using R, TFT, MM, and BR (see Figure 2 for results). We have mentioned that if N agent were not presented in such a situation, the population will quickly convergence to BR. When introduced even in small numbers, the Nash strategy survives by exploiting mainly random, then M and TFT (note that this situation does not conform to the trend of the ultimate winner being initially exploited by one of the strategies dominated by the initial winner). When only BR remains, the slight difference in favor of N for the head to head as for self play) makes the population converge to N.

Next, we perturbed the initial population in the second experiment by adding few learning agents using the FP strategy. The first interesting fact in Figure 3 is that initially the proportion of FP is increasing faster than the proportion of Nash. This is mainly due to the fact that FP exploits TFT agents much better than Nash does (see Table 3). After that, although Nash performs marginally better than FP against BR (gain of 0.0047 vs a loss of 0.036), FP exploits Nash and performs better in self play. Because of the presence of a learning strategy which more efficiently exploits the Nash strategy than the simple BR strategy, a more complex strategy, Nash, is ultimately eliminated. Once N is eliminated, we see a period of changing fortunes between BR and FP.

The population, left with two strategies, evolves to a dynamic equilibrium with a mix BR and FP strategies. This might appear incongruous with the head to head results since BR is winning against FP by 0.036. In self play, however, the FP agents are performing better (2.953 against 2.916). This is a very important factor, often determining the ultimate winners in evolutionary tournament. If a strategy exploits another, but fails to generate sufficient payoff playing itself, its performance will decrease as it becomes more numerous in the population. Thus it will fail to become dominant in the population. In Figure 4, we plotted the payoff obtained by FP and BR agents in a population consisting of only these two strategies in varying proportions. The head to head results provide the values used to compute the linear payoff functions, for an agent using strategy i : $f_i h_{ii} + (1 - f_i) h_{ij}$, where j is the other strategy in the population, f_i is the proportion of the population playing strategy i and h_{ij} is the (i, j) entry in Table 2. The lines intersects close to an equal

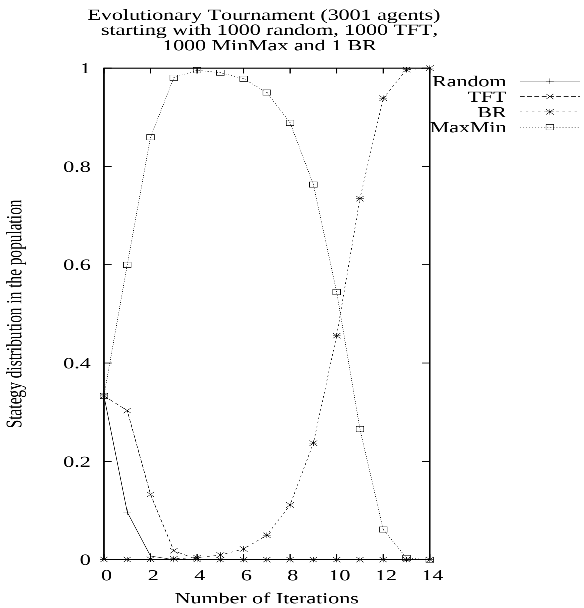

Figure 1: Evolutionary Tournament with 4 strategies (R, TFT, BR and MaxMin).

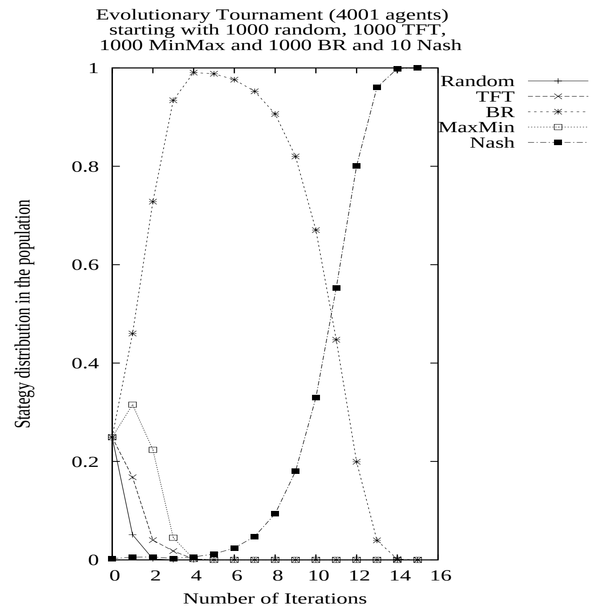

Figure 2: Evolutionary Tournament with 5 strategies (R, TFT, BR, MaxMin and Nash).

proportion of the two strategies in the population: this distribution would be a fixed point if sampling errors were not present. The actual selection mechanism used determines the convergence behavior of the population. The biased tournament selection mechanism we have used produces relatively large swings of the population proportions over successive generations that spans both sides of the fixed point proportion. A selection mechanism with less strong bias will converge either to the fixed point or to a dynamic equilibrium with narrower cycles.

Representative population We next consider a population containing all the strategies that are likely to be present in a large population: we have added to the 4 simple strategies (R, TFT, BR, M) the Nash strategy and the basic learning strategy FP. The result of the evolution, with uniform initial distribution of strategies, is presented in Figure 5: the population converges to a mixed strategy of FP and BR, as observed before in the perturbation of the simple strategy. It is interesting to notice that the proportion of Nash is first increasing, gaining over R. But then, as previously mentioned, FP and BR performs better than N. There are two important observations from this representative population:

- In a heterogeneous population a learning strategy like FP is preferable to the more commonly advocated Nash play.
- A relatively simple learning mechanism like BR can benefit from the presence of more sophisticated learning schemes like FP and outlive more complex strategies like Nash in the long run.

Next we perturbed this population by introducing one agent using the Saby strategy. The evolution is presented in Figure 6, and the population converges to a mix strategy distribution of Saby and BR. The proportion of Saby rise when there are no more Nash agents (Saby loses in head-to-head play against Nash). This time again, the introduction of a more sophisticated agent yield the disappearance of a complex strategy (FP) while a more simple strategy, BR, thrives. A similar analysis as in Figure 4 for populations of only Saby and BR agents show that the fixed point is for a proportion of $\approx 72.5\%$ of Saby agents.

Finally, we ran an experiment starting with a 1000 agents for R, TFT, BR, M, Nash, FP each, and we introduce one Saby agent and one BRFP agent. The outcome of the evolution (Figure 7) is a mixed strategy of BR, Saby and BRFP. BR is marginally present (around 1% of the population). The great majority of agents are using BRFP. From these last two results, we conclude that lesser known, learning players can grow to dominate the populations if agents adopt more successful strategies.

Conclusion and future work

We have evaluated several representative learning and non-learning strategies in a round-robin tournament format by playing two-player two-action iterative single stage games. The set of games used represents all the conflicting situations that can occur in a 2×2 game. The learning algorithms including fictitious play and a best response to it outperform non-learning players like the oft-quoted Nash player, which is a rational strategy for non-repeated games. Our results corroborate our hypothesis that evaluated over a large set of possible interaction scenarios, learning players not only have the potential, but do actually outperform non-learning players.

From our results it is clear that the learning players will

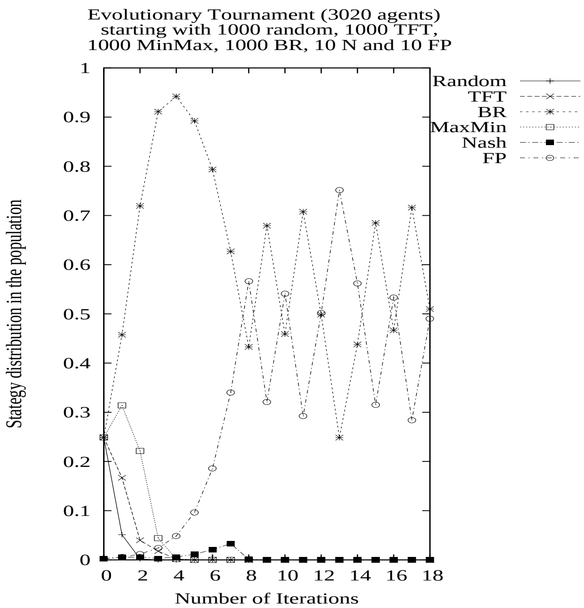

Figure 3: Evolutionary Tournament with 5 strategies (R, TFT, BR, MaxMin, Nash and FP).

typically outperform non-learning players when there is a variety of players in the tournament. We also notice that the learning players performed better in self play, an important consideration in evolutionary tournaments.

Head to head comparison of strategies enables us to study the evolution of the strategy distribution in a potentially large population of agents. The selection mechanism we used is a variant of the tournament selection. We developed a finite-population model to compute analytically the strategy distribution of the next generation. The analytical model gives us the ability to calculate the eventual population distribution given starting distribution of strategies without having to run costly, time-consuming experiments.

This model was used to study the evolution of likely populations of agents: a population of agents using simple strategies and a population of agents using representative strategies including a more complex learning strategy and playing a Nash equilibrium. The results indicate that the outcome of the evolution is dependent upon the initial strategy distribution. It is interesting to notice that a population is able to adopt a more sophisticated strategy, even though initially used by a minority of the agents.

The assumption that any agent can observe the strategy of any other agent at the end of the round robin tournament may not be realistic. One agent may only reveal its strategy to a small number of other agents instead of publishing it to the entire society. We are planning to study this social network effect on the evolution of the strategy distribution. We also plan to study the effects of other selection schemes on the population dynamics.

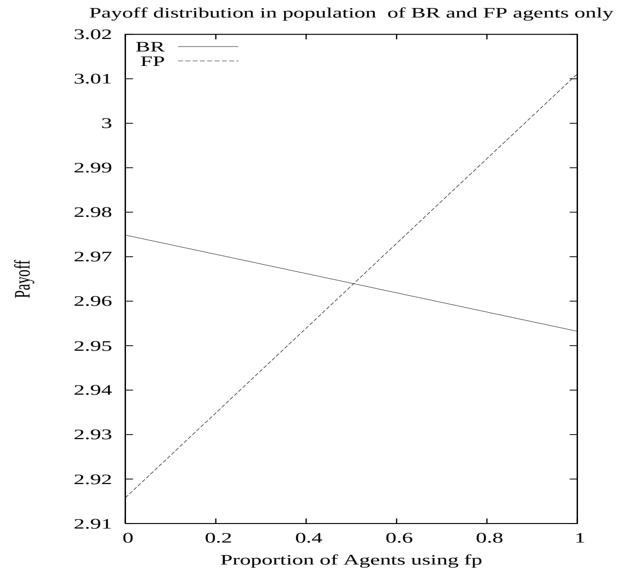

Figure 4: Payoff of FP and BR function of the proportion of FP agents.

Acknowledgments This work has been supported in part by an NSF award IIS-0209208.

References

Airiau, S., and Sen, S. 2003. Tournament-based comparison of learning and non-leanring strategies in iterated single-stage games. In *Proceedings of the 5th workshop on Decision Theoretic and Game Theoretic Agents*, 1–6.

Bowling, M., and Veloso, M. 2001. Rational and convergent learning in stochastic games. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, 1021–1026.

Brams, S. J. 1994. *Theory of Moves*. Cambridge University Press, Cambridge: UK.

Claus, C., and Boutilier, C. 1998. The dynamics of reinforcement learning in cooperative multiagent systems. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence*, 746–752. Menlo Park, CA: AAAI Press/MIT Press.

Deb, K., and Goldberg, D. 1991. A comparative analysis of selection schemes used in genetic algorithms. In Rawlins, G. J., ed., *Foundations of Genetic Algorithms*, 69–93. San Mateo, CA: Morgan Kaufman.

Fudenberg, D., and Levine, K. 1998. *The Theory of Learning in Games*. Cambridge, MA: MIT Press.

Littman, M. L. 1994. Markov games as a framework for multi-agent reinforcement learning. In *Proceedings of the Eleventh International Conference on Machine Learning*, 157–163. San Mateo, CA: Morgan Kaufmann.

Mundhe, M., and Sen, S. 2000. Evaluating concurrent reinforcement learners. In *Proceedings of Fourth Interna-*

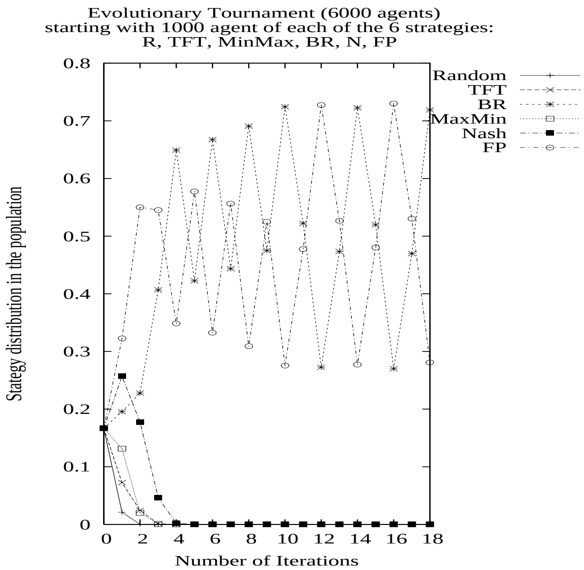

Figure 5: Evolutionary Tournament with 6 strategies (R, TFT, BR, N, M, FP).

tional Conference on MultiAgent Systems, 421–422. Los Alamitos, CA: IEEE Computer Society.

Nash, J. F. 1951. Non-cooperative games. *Annals of Mathematics* 54:286 – 295.

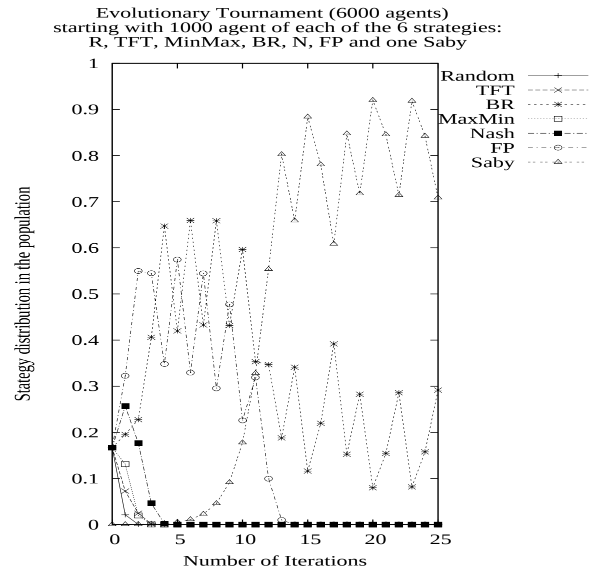

Figure 6: Evolutionary Tournament with 7 strategies (R, TFT, BR, N, M, FP and Saby).

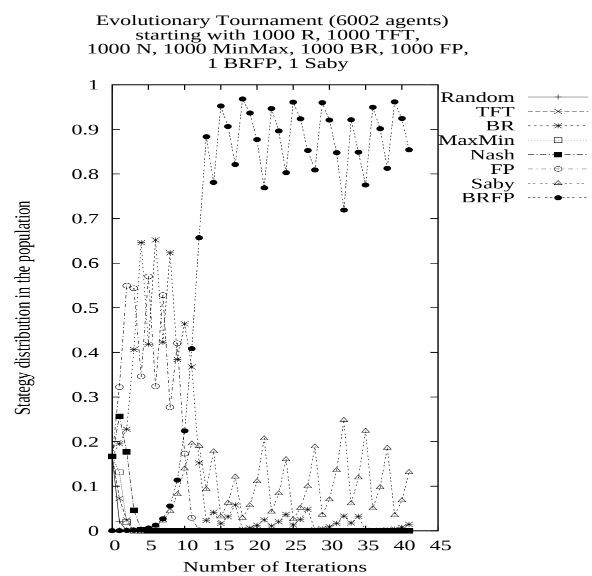

Figure 7: Evolutionary Tournament with 8 strategies (R, TFT, BR, N, M, FP, BRFP and Saby).