Evolutionary tournament-based comparison of learning and non-learning strategies for iterated games

Stéphane Airiau and **Sandip Sen** and **Sabyasachi Saha**
The University of Tulsa
Mathematical and Computer Sciences Department
600 South College Avenue
Tulsa, OK 74104
{stephane,sandip,saby}@utulsa.edu

Abstract

Evolutionary tournaments have been used as a tool for comparing 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 that embody some of the above approaches on an experimental testbed of all possible structurally distinct 2x2 conflicted games with ordinal payoffs. This set of bimatrices provides a baseline, neutral testbed for comparing strategies. We discuss the testbed, our choice of representative learning and non-learning strategies and relative rankings of these strategies ranked by cumulative score in tournaments. We also study the performance of the strategies in an evolutionary tournament. Finally, we provide some analysis of the observed results to highlight the advantage of learning strategies.

Introduction

Learning and reasoning in single or multistage games have been an active area of research in multiagent systems (Banerjee, Sen, & Peng 2001; Bowling & Veloso 2001; Claus & Boutilier 1998; Hu & Wellman 1998; Littman 1994; 2001; Littman & Stone 2001). 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 the 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. We will refer to a player who always plays a Nash strategy as a Nash player. Though it follows from 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.

In this paper, our goal is to evaluate relative performances of representative learning and non-learning approaches on a neutral but extensive testbed. For the testbed we adopted the set of all possible structurally distinct 2x2 games with ordinal payoffs without a single preferred outcome.

We compare the representative strategies in a tournament format with each player (corresponds to a strategy) playing every other player for all the 57 games both as the row and the column player. For a given game, n iterations are played which gives the learning players an opportunity to approximate the strategy being used by its opponent. The games are played under the assumption of complete information, i.e., each player is aware of both its own and the other player's payoff matrix. Cumulative payoffs are used to order the performance of the players. A one shot tournament may not be representative of the strength of a strategy. Hence, we study the evolution of a large population where all the strategies are equally present at the beginning.

In the following we present the testbed with a characterization of the set of games used. Next, we introduce the strategies used by the different players, the results from the basic tournament, an evolutionary tournament framework with corresponding results, and summary observation of the results.

Tournament structure

In this section, we describe the basic tournament structure given the set of matrices and a selection of strategies. First we present the matrices that are used and then the format and ranking procedure.

Testbed: Structurally distinct 2x2 conflicted games with ordinal payoffs

We only consider a subset of the possible 2x2 payoff matrices where agents have a total preference order over the four possible states. We will use the numbers 1, 2, 3, 4, as the preference of an agent for a state in the 2x2 matrix, with 4 being the most preferred. Though the payoff represent ordinal payoff, we treat them as cardinal payoff.

This set of 57 matrices represent all the distinct conflicting situation with ordinal payoffs(Brams 1994). We believe

this is an interesting set on which we can test the performance of strategies or learning algorithms. If one strategy is considered to be robust, it must be the case that overall, competing against a variety of opponents, this strategy manages to perform well enough. It might be the case that it cannot perform well against some opponents in some particular games, some strategies might be more suitable in some situation than others. However on the average, a robust solution should do well. Hence, if we run a tournament between different strategies, we might expect that the winner of such tournament present a practical solution in a general problem. Also, in the context of a large population, a strategy might be well suited to exploit some strategies, and be weak against others, especially in the case where the proportion of the different strategies are not equal. Hence, it is also important to test the performance of the strategies in evolutionary tournaments.

Tournament setting

No two matrices, in the chosen set of 57 games, represent the same game if we rename the players, or rename their actions. The construction of the matrices, however, introduces a bias: the player who is considered to be the column player has an advantage. For instance, if we study the payoffs of the pure Nash equilibria, in 31 cases, the payoff of the column player is strictly higher than the payoff of the row player, in 9 cases they are equal, and in 2 cases, the payoff is greater for the row player.

To be fair, each player plays every other player both as a column and as a row player for each of the matrices. Each game is iterated 100 times. Because the action space is small, we assumed that 100 iterations are reasonable for players to adapt their strategies when using a learning approach. In order not to penalize the learning players that are likely to perform badly during the early rounds, we accumulate the payoffs of the players only over the last 50 iterations of the game. In this paper, each player has complete information over the game. Each player is provided both its own and the other's payoff matrices. The players are not allowed any other means of communication apart from expressing their action at each iteration.

To summarize, for each pairing of two players, a player will play 100 iterations as a row player, and 100 iterations as a column player on each of the 57 game matrices. The score of a player is computed by accumulating the payoffs of the last 50 iterations for each game, hence over $57 * 2 * 50 = 5700$ decisions, for a particular opponent.

Players

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.

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

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

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.

Results

In the following, we summarize the results of the tournament in two different format: ranking based on cumulative scores and head to head comparisons.

Tournament Results

In the tournament we ran, each strategy introduced in the previous section is represented by one player. As shown in Table 1, the player using the strategy of Saby wins the tournament, followed closely by a player using best response to fictitious play. In the table, we show the average score per iteration to show the average level of preferences obtained. All but the random player obtain close to their second most preferred choice on the average.

rank	player	Average score per game
1	Saby	3.00
2	BRFP	2.99
3	FP	2.98
4	N	2.94
5	BR	2.93
6	MaxiMin	2.79
7	TFT	2.75
7	R	2.41

Table 1: Strategy ranking based on tournament.

BRFP, Saby's player and FP are learning approaches that model their opponent's behavior, and they take the three first places. The Nash player, models its opponent as a rational player, and it comes in at the fourth position. It was surprising to see that best response to last move, a fairly straightforward strategy, performs almost at the same level as the Nash player. The tit for tat player, which was a good choice in a tournament setting involving only the prisoner's dilemma game turns in a relatively disappointing performance, and finishes ahead of only the random player.

Two facts caught our attention in this results. First, 3 learning players perform better than the Nash player. It might be the case that the other players are taking advantage of the randomness introduced in the Nash player (in the cases where there are multiple Nash equilibria, or in the case when the Nash equilibrium is a mixed strategy). It might also be the case that learning players play better against subrational strategies like BR. Another point of interest was the top ranking of BRFP: one can expect that the performance of the best response to fictitious play outperform fictitious play. However, since the other players are not using fictitious play, it is not clear how well BRFP will do against them. One possible explanation is that BRFP may be exploiting the FP to obtain a large payoff, and performing satisfactorily against the others. In order to understand better this overall result, we looked closely into the head to head results which we present in the next section.

Head to Head Comparisons

In this section, we describe head to head results over the 57 games, derived from Table 2. This table contains the average score obtained by the strategy of the row while playing against the strategy of the column over 100 instances of the tournament. These results have a small standard deviation (see Tabletab:varh2h) due to the use of learning and random players. From the head to head results, we can compute the net difference of score of the player (substracting from Table 2 its transpose). for lack of space, we did not present

this table. The observation of the head to head results leads to the following conclusions:

Nash: Its main gain comes from its games with the random player and the MaxiMin player (a net win of respectively 0.53 and 0.16, the other result not being significant).

FP: It is the best strategy if the opponent is a R, TFT or N player. It loses noticeably against BRFP and Saby's players (a net loss of resp. 0.27 and 0.15). The game against the random player plays an important role in the ranking of this player (with a net win of 0.61. Also, without BRFP, in the tournament, its average accumulation would have been much better, and it would win the tournament.

BRFP: It is the best strategy if the opponent is a BR, FP or Saby's player. It loses narrowly against only TFT and N, but the results are not significant. This is expected since these two strategies are completely different from FP. But it makes up by obtaining high gains against BR (which is a degenerate version of FP with a single memory cell), FP (for which it is designed to play optimally) and Saby's (resp. a net win of 0.20, 0.28 and 0.17). Moreover, it is the only player which is able to defeat significantly Saby's player head to head.

Saby's player is the best strategy (followed by Nash) against BRFP.

Also, while considering the diagonal of the matrix, we can compare the self play results. Saby's strategy perform the best with a score of 2.99, followed by fictitious play with a score of 2.955.

Evolutionary Tournament

In the preceding instances of the tournament we ran, each strategy is represented by one player. The player has a fixed strategy and we compare the average payoffs for a player with a given strategy. We have also investigated the effect of an evolutionary mechanism upon the demography in terms of strategy distribution in the population in a tournament with multiple players per strategy. We wanted to find out the existence and if so the nature of the equilibrium strategy distribution of the population and the rate of convergence to that distribution starting from uniform strategy distributions.

An equilibrium refers to convergence to a stable state (population distribution) or a set of states. Payoffs for a given strategy can change with the changing distribution (or demography) of the population. It is possible that a given strategy receives high payoff only in the presence of some other strategy or that two strategies are mutually reinforcing, leading to an equilibrium distribution. This can produce interesting strategy distribution dynamics as the population evolves.

A round robin tournament is played for each generation. In each generation, each player plays each other player in every matrix and both as a row and a column player. The resultant cumulative score represents the performance of the player in that generation of the tournament. We are more interesting in the dynamics of the evolution than a fat convergence to an equilibrium. Hence, we want to test different selection mechanism and study whether they give rise to the same kind of dynamics. To select the strategy distribution in the population in the next generation, we considered three different schemes(Deb & Goldberg 1991): the pure and a modification of tournament selection, and fitness proportionate selection. The modification of the tournament selection (see Algorithm 1) adds a flavor of the fitness proportionate selection mechanism. Tournament selection only need local information, i.e. the score of two players chosen randomly. The modification tends to increase the likelihood of choosing players which performs better. The pure tournament selection selects the two strategies ρ_0 and ρ_1 from a uniform distribution whereas the modified scheme selects them with a probability proportionate to their score. A priori, the convergence using the modified version of the tournament selection should be faster.

Algorithm 1 Modified 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 and ρ_1 according to $Prob$
    $newstrat(k) \leftarrow strat(argmax_{i\in\{0,1\}}(score(\rho_i)))$
  for every player k **do**
    $strat(k) \leftarrow newstrat(k)$

We run tournaments starting with a population where the strategies are uniformly distributed, i.e., all strategies have equal representation. In order to avoid to run the games each time, we simulate the tournament. For a game between two player p_1 and p_2, the score obtained by $p_{i\in\{1,2\}}$, is a sample drawn from a normal distribution of the corresponding mean and variance from tables 2 and **??**. This is only an approximation of the tournament since we draw the score from independent distribution, which is not correct. However, we did not observe different trend when we compare the actual tournament on small settings. Hence we believe this is a good approximation and the variance of the results do not play an important role in the results.

Figures1, 2 and 3 present the result of a simulation ran with a initial population of 100 agents for each strategy, using the three different selection mechanism. The three systems converge to a mixed population of agents using Saby and BRFP strategy. When these two strategies are present in the system, there is an equilibrium point for a proportion of 0.2017% of agents using Saby and 0.7983% of agents using BRFP. At this point, all the agents have the same score as shown in Figure 4. It does not come as a surprise that these two strategies do well, however, it is surprising to observe a mix equilibrium.

It is also interesting to note that the three selection mechanisms have a similar dynamics, though on different scale. At the beginning, as expected from the results of the tournament, saby, BRFP, FP and FP increases their proportion.

	RN	TFT	N	BR	FP	BRFP	MaxiMin	Saby
RN	2.5003	2.4994	2.3692	2.4087	2.3869	2.3680	2.3138	2.3990
TFT	2.5014	2.5198	2.9273	2.8216	2.7507	2.8863	2.5824	2.9724
N	2.9036	2.9283	2.9399	2.9337	2.9267	2.9444	2.9036	2.9393
BR	2.9161	2.7404	2.9360	2.9056	3.0019	2.9339	3.0823	3.0158
FP	2.9944	3.0107	2.9429	2.9928	2.9551	2.9047	3.0732	2.9427
BRFP	2.8826	2.8598	2.9175	3.1305	3.1823	2.9174	2.8784	3.1158
MaxiMin	2.9980	3.1374	2.7401	2.7056	2.7156	2.7354	2.6316	2.6952
Saby	2.9144	3.0140	2.9336	3.0338	3.0915	2.9491	3.0791	2.9904

Table 2: Head to head results - score obtained by the row player while playing against the column player

	RN	TFT	N	BR	FP	BRFP	MaxiMin	Saby
RN	0.016032	0.015864	0.012480	0.013441	0.030937	0.011822	0.009820	0.014533
TFT	0.012642	0.071483	0.007061	0.033716	0.043553	0.001779	0.001818	0.018885
N	0.011537	0.008352	0.006767	0.008595	0.025941	0.006234	0.005641	0.008282
BR	0.015559	0.001810	0.009009	0.000000	0.000000	0.000000	0.000000	0.000478
FP	0.014098	0.042547	0.007501	0.000000	0.000000	0.000000	0.000000	0.008326
BRFP	0.012979	0.038393	0.006967	0.000000	0.000000	0.000000	0.000000	0.002516
MaxiMin	0.011796	0.001845	0.007582	0.000000	0.000000	0.000000	0.000000	0.003178
Saby	0.013698	0.023523	0.007718	0.008456	0.004259	0.002609	0.000229	0.014915

Table 3: Head to head results - standard deviation of the score obtained by the row player while playing against the column player

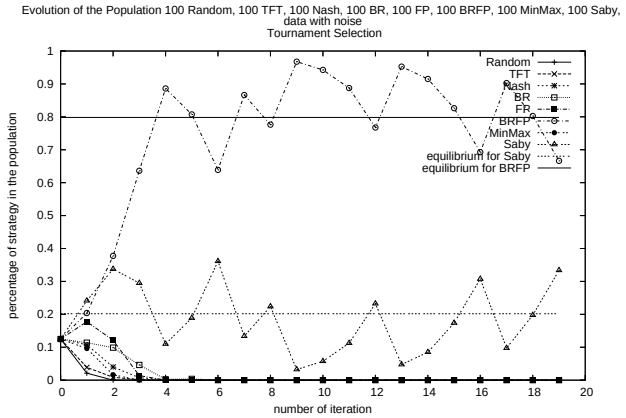

Figure 1: Evolutionary Tournament with 100 agents for each of the 8 strategies. Tournament Selection

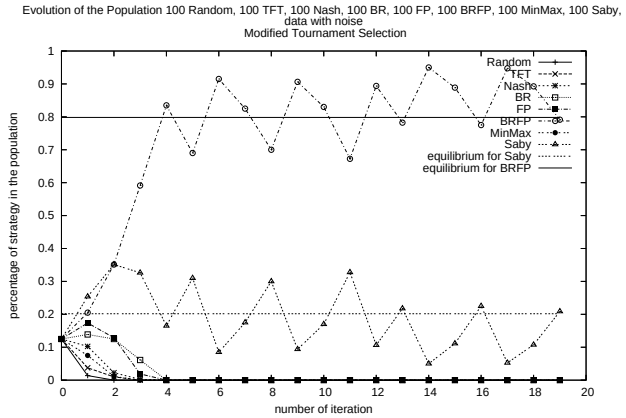

Figure 2: Evolutionary Tournament with 100 agents for each of the 8 strategies - Modified Tournament Selection.

Then, because of the presence of more agents using FP, BRFP can exploit them better, hence we see that the proportion of FP agent decreases, and the proportion of BRFP improves faster than the proportion of Saby. At this point, most of the agents are using either BRFP and Saby, and the proportion starts to get closer to the equilibrium point.

Tournament selection and the modified tournament selection needs 6 iterations to reach their equilibrium, whereas the fitness proportionate mechanism reaches convergence after $\approx$ 180 iterations. Since the convergence is fast, the modification of tournament selection does not play an important role in this problem. However, the selections mechanism for the tournament selection and its modification makes the system bounce back and forth the equilibrium point. Experiments using larger numbers of evolutions shows that the system keeps on oscillating without dampening, and we also observed these oscillations in larger systems (the same phenomenon occurs with a population initially containing 1000 agents for each strategy).

Conclusion

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.

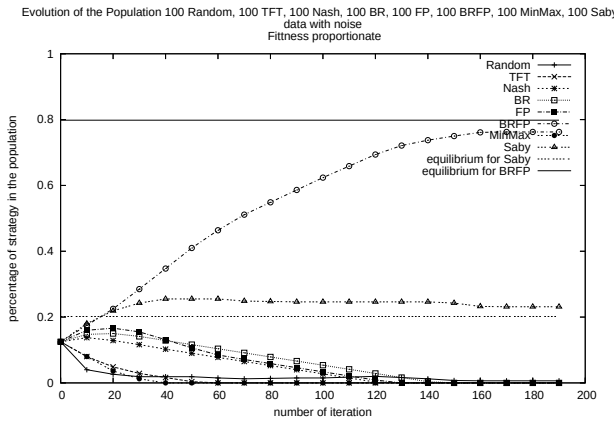

Figure 3: Evolutionary Tournament with 100 agents for each of the 8 strategies - Fitness proportionate Selection.

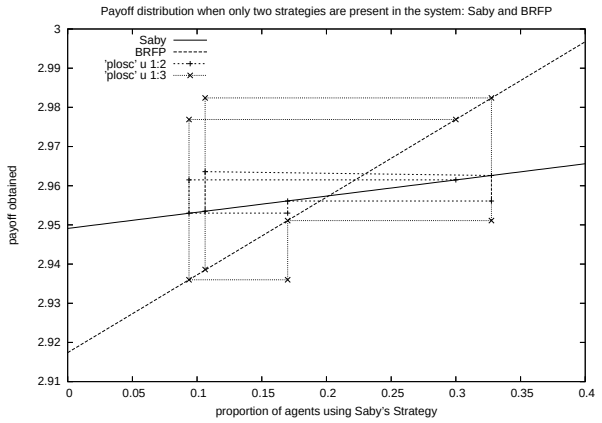

Figure 4: Payoff Vs proportion of agents using Saby in a system with only Saby and BRFP - Equilibrium point at 0.2017

The evaluation set consists of all structurally distinct 2x2 conflicted ordinal-preference games and provides a fair and extensive set of testing scenarios.

The learning algorithms including fictitious play and a best response to it outperform players like Nash. Though the actual ranking is dependent on the exact set of players involved, it can be argued that the learning players will 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. It also came as a surprise that averaged over all the games, the Nash player could significantly outperform only the random player. Also the fictitious play player loses out to the player who plays best response to it, and can only fare well by outperforming the random player. The evolutionary tournament reinforced the results from the basic, single-stage tournament.

While the current set of experiments are run over all structurally distinct conflicted 2x2 games with ordinal payoff, it would be interesting to see if these results generalize to large samples of randomly generated $n \times n$ games with cardinal payoff. Finally, we observed the evolution of population using different initial strategy distributions. In some cases, few good agents can take over an entire population, especially in the case of tournament selection. We are planning to investigate further these effects and the importance of the selection mechanism.

References

Banerjee, B.; Sen, S.; and Peng, J. 2001. Fast concurrent reinforcement learners. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, 825–830.

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.

Hu, J., and Wellman, M. P. 1998. Multiagent reinforcement learning: Theoretical framework and an algorithm. In Shavlik, J., ed., *Proceedings of the Fifteenth International Conference on Machine Learning*, 242–250. San Francisco, CA: Morgan Kaufmann.

Littman, M. L., and Stone, P. 2001. Implicit negotiation in repeated games. In *Intelligent Agents VIII: AGENT THEORIES, ARCHITECTURE, AND LANGUAGES*, 393–404.

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.

Littman, M. L. 2001. Friend-or-foe q-learning in general-sum games. In *Proceedings of the Eighteenth International Conference on Machine Learning*, 322–328. San Francisco: CA: Morgan Kaufmann.

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