Available online at www.sciencedirect.com
ScienceDirect
Artificial Intelligence 171 (2007) 423–428
Artificial Intelligence
www.elsevier.com/locate/artint

Multi-agent learning and the descriptive value of simple models

Ido Erev [a,*], Alvin E. Roth [b]

[a] *Technion, Israel*
[b] *Harvard University, USA*

Received 25 April 2006; received in revised form 20 October 2006; accepted 22 November 2006

Available online 26 January 2007

Abstract

Behavioral research suggests that human learning in some multi-agent systems can be predicted with surprisingly simple "foresight-free" models. The current note discusses the implications of this research, and its relationship to the observation that social interactions tend to complicate learning.

Keywords: Reinforcement learning; Fictitious play; Equivalent Number of Observations (ENO); Reciprocation

1. Introduction

Shoham, Powers, Grenager [28] ask: "if multi-agent learning is the answer, what is the question?" Their search for interesting questions focuses on the observation that the analysis of learning in multi-agent settings tends to be more complex than the analysis of individual learning. For example, the move from one-person to two-person games opens the door to phenomena like the role of teaching and reciprocation, as well as learning.

In this comment we try to extend the search for interesting questions by considering the possibility that an increase in the number of agents can simplify the task of capturing human learning. The basic idea is that in many multi-agent interactions, the expected effect of each agent on the environment and on other agents may be very small. In addition, the opportunity to observe the behavior of other agents reduces the importance of exploration.

Multi-agent transportation games are a natural example. Each agent in these games has to select a transportation route. The aggregate choices determine the payoffs, but the effect of each agent is typically small. And when the agents can observe each other, exploration is less important. Each agent can learn from the experience of similar agents that he or she can observe.

Our research suggests that in these and similar games, multi-agent learning can be usefully described and predicted with surprisingly simple "foresight-free" reinforcement learning models; models that ignore the possibility that current choices change the environment, and/or provide information needed to maximize long term outcomes. The fictitious play rule [25] is an example of a foresight-free learning model, in which at trial t, an agent selects the action that has led to the best average outcomes in the first $t-1$ trials. A stochastic variant of this model is considered below.

In Section 2 we review some research that will help illustrate the value of simple learning models in describing human behavior in the lab. In Section 3 we return to Shoham et al.'s question about what kinds of questions these

* Corresponding author.
E-mail addresses: erev@tx.technion.ac.il (I. Erev), aroth@hbs.edu (A.E. Roth).

0004-3702/$ – see front matter
doi:10.1016/j.artint.2007.01.001

results make us optimistic that we can answer. We formulate three questions that the study of multi-agent learning might help answer, involving the speed at which learning converges to equilibrium in different strategic environments, individual decision making in one-shot tasks, and behavior in natural settings.

2. The value of simple models of human learning in multi-agent settings

We focus on the "descriptive agenda" (see [28]), and consider how learning models capture human behavior observed in the laboratory. As suggested by Shoham et al., an increase in the number of agents increases the set of feasible strategies, and can complicate models of learning. Indeed, experimental research suggests that agents consider complex reciprocation-related strategies in two-person games (see [24]). Nevertheless, we argue that when the number of agents is large, simple descriptive models of individual learning can sometimes predict multi-agent learning surprisingly well.

We do not argue that multi-agent games have simpler equilibria, or that behavior in all large multi-agent games can be captured with simple models. But the preliminary evidence suggests that there are many interesting and important multi-agent environments in which behavior can be usefully approximated with simple models. Three observations that support this assertion are presented below.

2.1. Reciprocation in two- and four-agent settings

In their classic book, Rapoport and Chammah [24] demonstrated that most human agents learn to reciprocate in certain repeated two-person prisoner dilemma games. In one of their experiments, participants played the game in Table 1 for 300 trials against the same opponent with immediate feedback after each trial (and without knowing how many trials would be played). The results showed a low initial rate of cooperation (about 30% in the second block of 50 trials), and an increase in cooperation with experience. The cooperation rate in the last block was higher than 50%. It is easy to see that these results cannot be captured with a simple learning model that ignores teaching and/or reciprocation (see [14]). Such models predict a *decrease* in cooperation over time.

To evaluate this failure of basic learning models, Danieli [8] ran two versions of the experiment described above. The first was a computerized replication of the original study. The participants were run in cohorts of four that were divided into two pairs. Each pair interacted 300 times. The results of this condition were very similar to the original results. The proportion of cooperation in the last block of 50 trials was 65%.

The second condition was identical to the first with the exception that the four participants in each cohort were randomly re-matched after each trial. This change had a dramatic effect on the results. The proportion of cooperation in the last block of 50 trials dropped to 15%.

These results suggest that simple foresight-free models that fail to capture 2-person interaction can already do a much better job at capturing 4-person interaction. The results suggest that the effect of the factors ignored by these models (like reciprocation and teaching) quickly drop when the number of agents increases.

We note that not only the number of players is important, but also the pattern of interaction. When (even many) players are paired in fixed pairings for even a few periods, reciprocation and cooperation can develop. For example, Bereby-Meyer and Roth [4] report an experiment one of whose conditions reproduces the effect earlier observed by Selten and Stoecker [27] and Andreoni and Miller [1]. When subjects are rematched to play with different counterparts in a 10-period repeated prisoner's dilemma, in which the number of periods to be played with the same partner is common knowledge, the players can learn with experience to cooperate with high probability in the first few periods, and to cooperate with low probability in the final periods of each repeated game. (Bereby-Meyer and Roth also show how environmental factors that affect the speed of learning can also have large consequences for how much cooperation is achieved, a point to which we will return later.)

Table 1
One of the prisoner dilemma games studied by Rapoport and Chammah [24]

Player 1	Player 2	
	C	D
C	(1, 1)	(−10, 10)
D	(10, −10)	(−1, −1)

2.2. The role of exploration

Evaluation of the descriptive value of foresight-free learning models in individual choice tasks [11] reveals high sensitivity to feedback. Foresight-free models fit the observed learning well when the environment is static and the feedback includes information concerning both the payoffs actually obtained and the foregone payoffs (that would have been achieved from actions that weren't chosen). However, models of this type tend to fail when the feedback is limited to the obtained payoffs. When the feedback is limited, simple foresight-free models tend to predict a strong "hot stove" effect (learning to prefer the safer alternative, see [9]). Human behavior exhibits a much weaker effect of this type.

This failure of simple foresight-free models is likely related to the exploration of alternative actions. Exploration has little effect when foregone payoffs are known, but can drive learning when the feedback is limited. Modeling human exploration exploratory behavior is not trivial. It seems that humans can learn to increase the exploration rate when exploration is likely to increase long term expected return. Thus, it seems that the decision to explore involves some foresight.

This shortcoming of foresight-free learning models appears to be less relevant in the context of multi-agent games in which others' actions are observable (see related arguments in [5,7]). When agents can observe the behavior and the outcomes of similar agents, these observations provide information concerning the foregone payoffs. For example, in many multi-agent transportation games foregone payoffs can be inferred from the outcomes obtained by other agents.

2.3. Generality and predictive value

In Erev and Roth [13] we examined if basic foresight-free reinforcement learning models that ignore the dynamic features of the environment can capture behavior in games with unique mixed strategy equilibrium. The data considered in that paper included results of published experiments that show apparently complex patterns. In some of the games studied, experience quickly moved behavior towards equilibrium [23], while in other games there was little correspondence between the observed results and equilibrium [29]. Moreover, many of the studies show non-monotonic learning trends. The left-hand side of Fig. 1 presents five of the twelve games analyzed in that paper. The experimental results and equilibrium predictions are summarized in the left-hand column by the proportion of "A" choices.

The right-hand columns show the results of computer simulations in which virtual agents that behave according to simple foresight-free learning models play each of the games. The virtual experiments were run under the same conditions as the original studies. In particular, they included the same number of trials as the original studies. The results show that all three models presented in Fig. 1 capture the main behavioral trends. Good fit between the models and observed behavior was obtained both when the observed behavior is near and far from equilibrium predictions.

It is important to emphasize that the high correspondence between the observed human behavior and the models is not a result of overfitting the data. Similarly good correspondence was observed when the parameters of the learning models are estimated on one set of games, and used to predict behavior in a second set of games. A clear demonstration of this point is provided in [15]. That paper uses a stochastic variant of the fictitious play model (see [16,17]) to predict behavior in ten randomly selected constant sum games. The probability of selecting alternative k at trial t is modeled as:

$$P_k(t) = \frac{e^{q_k(t)\lambda/s(t)}}{\sum_{j=1}^{2} e^{q_j(t)\lambda/s(t)}} \tag{1}$$

where $q_j(t)$ is the propensity to select strategy j, λ is a payoff sensitivity parameter, and $S(t)$ is a measure of experienced regret.

The adjusted propensity to select alternative j at trial $t+1$ is:

$$q_j(t+1) = (1-w)q_j(t) + w \cdot x_j(t) \tag{2}$$

where $x_j(t)$ is the payoff of j in trial t, and w is a parameter that determines the weight of this payoff. The initial value is $q_j(1) = 0$.

The level of experienced regret, $S(t)$, is modeled as the weighted average of the difference between the obtained and the maximal payoff:

$$S(t+1) = (1-w)S(t) + w\left|\max(t) - x_j(t)\right| \tag{3}$$

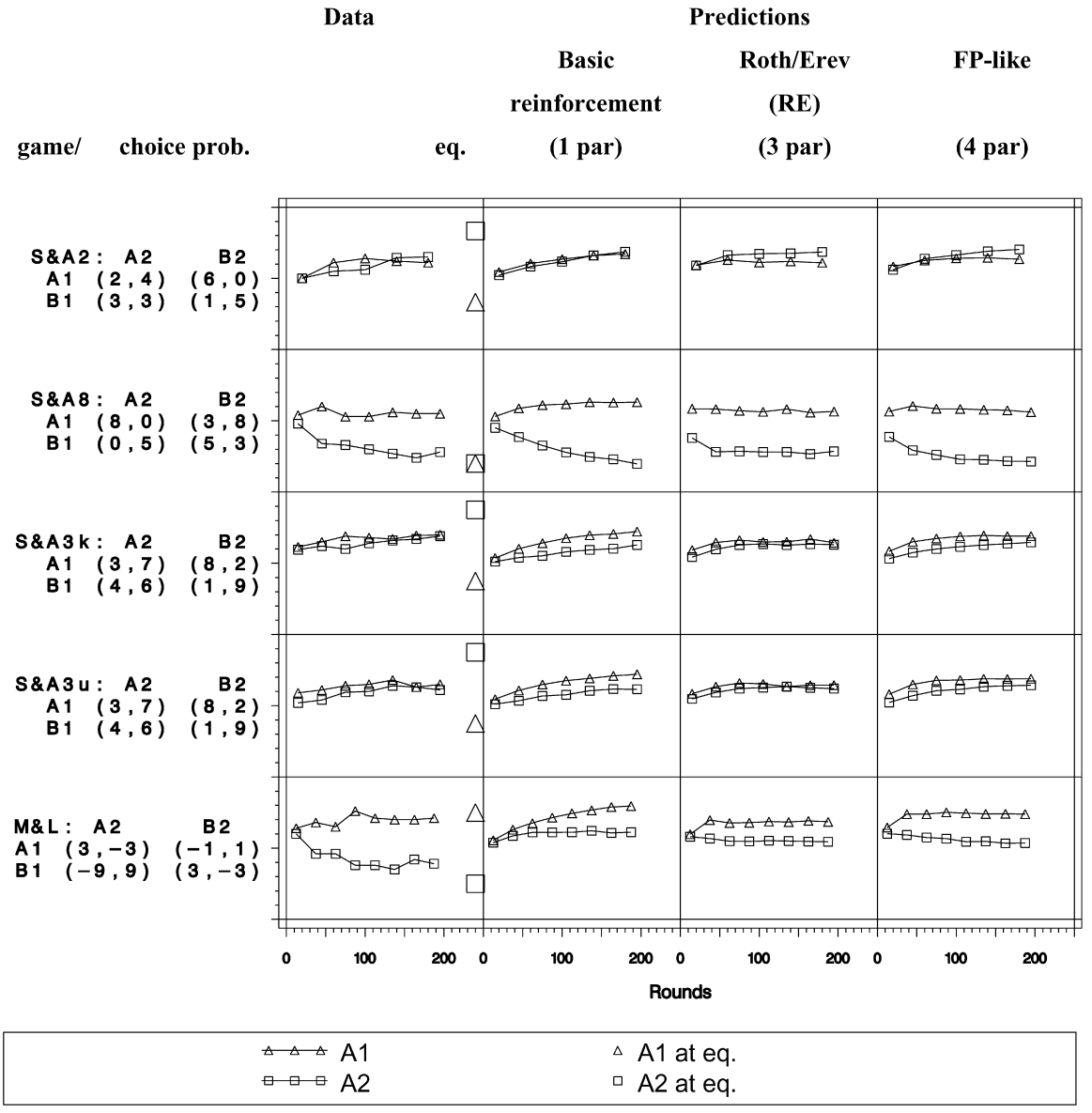

Fig. 1. Repeated 2 × 2 games [21,29]. In the top four games each payoff unit increases the probability of winning (by 1/6 in S&A2, by 1/8 in S&A8, and by 1/10 in S&A3k and S&A3u). In M&L payoffs were directly converted to money. Each cell in the left-hand column presents the experimental results: The proportion of A choices over subjects in each role (grouped in 5 to 8 blocks) as a function of time (200–210) trials in all cases. The three right-hand columns present the models' predictions in the same format. The equilibrium predictions are presented at the right-hand side of the data cells. (Adapted from [13].)

where $\max(t)$ is the maximal payoff obtained in trial t over the 2 alternatives. The initial regret level is set to equal $S(1) = \lambda$. The parameters of the models were estimated by Ert and Erev's [16] in a study of individual decisions. The estimated value are $w = 0.45$, and $\lambda = 2.7$.

The value of the model's predictions was assessed by computing the model's *equivalent number of observations* (ENO). A model's ENO is an estimate of the expected size of the experiment that has to be run so that the mean observations are more accurate than the models' predictions. The estimated ENO of the learning model was 16.7. That is, the model's prediction of how subjects would behave in a given game is more accurate than the prediction based on the average behavior of 16 other pairs of subjects observed playing the same game. In comparison the estimated ENO of the equilibrium prediction is 1.17. That is, equilibrium provides a better prediction of how a pair of

players will play the game than the observation of a single pair of other players playing the game, but not as good a prediction as can be made by observing two pairs of players.

Additional studies show that the good predictive value of simple learning models is not limited to two-person games with unique mixed strategy equilibrium. Similar results were observed for extensive form games [26]; Coordination games [12], and team games [6]. In those papers, learning models were able to predict human behavior both in games in which observed behavior converged quickly to equilibrium, and in games in which behavior remained persistently far from equilibrium.

3. Three questions

Research such as that reviewed above suggests that the study of learning in multi-agent systems can shed light on some important questions.

One question is, in what environments does human learning converge to equilibrium behavior in the intermediate term? (That is, should we expect to see learning in multiplayer economic environments proceed on a time scale that would allow equilibration to occur faster than big changes in an economic environment? Or fast enough to observe equilibration in the lab? see cf. [26].)

The model presented above and the data it summarizes suggest that the answer is "no" in at least some important cases, and "yes" in others.

One environmental factor that robustly slows learning in games is payoff variability (see [4,18,22]). This is consistent with the observation from the psychology literature that partial reinforcement (adding randomness to the link between an action and its consequences while holding expected payoffs constant) slows learning. This effect can be considerably magnified in multiplayer games such as the repeated prisoner's dilemma. When others are slow to learn to cooperate, the benefits of cooperation are reduced, which further hampers cooperation. That is, in a game, what there is to learn depends at least in part on what others are learning. Bereby-Meyer and Roth observed that a small change in the payoff environment, which changes the speed of individual learning, can thus have a large effect on collective behavior. These results suggest that there may be interesting comparative dynamics that can be derived by paying attention to the fact that at least some behavior is learned from experience.

Closely related to this is the question of whether learned decisions, i.e. decisions made from experience, are similar to decisions that are made based on a complete description of the incentive structure (like the decisions studied in Kahneman and Tversky's [19] classic work). The answer again is, at least sometimes, no.

Indeed, some of the deviations from rational choice observed in decisions made from experience are in the opposite direction of the deviation from rational choice captured by prospect theory. In particular, prospect theory implies oversensitivity to low probability outcomes, and decisions from experience reflect under sensitivity to these events (see [3]). The high estimated weighting parameters ($w = 0.45$) reflect high sensitivity to recent outcomes. Therefore, rare events that are not likely to occur recently receive, in many cases, little attention.

A third question involves the practical implications of the behavioral research considered here. Specifically, can the study of human learning in the lab shed light on human behavior in natural settings? We believe that the answer to this question is yes. In this respect we are encouraged by the robustness of the experimental results. The fact that a wide set of experimental conditions can be captured with a simple 2-parameter model suggests that the observed behavior reflects general behavioral tendencies. We are similarly encouraged by observations of interesting empirical phenomena that reflect underweighting of rare events and the payoff variability effect (see review in [10]). An important example is the "it won't happen to me" phenomenon: the observation that people tend to violate safety rules even when this behavior impairs their expected utility. Finally, when we are able to compare laboratory results with the behavior that we observe market participants learning in field environments (see e.g. [2,20]), what we see in the lab substantially reproduces what we see in the field.

4. Summary

Most studies of multi-agent learning focus on fact that the social interaction complicates the learning task. Here we highlight the fact that in certain settings an increase in the number of agents reduces the sensitivity of the learning process to exploration and environment-modifying strategies. (One way to think about this is by analogy to environments in which each player has a negligible effect on the environment, so that people may behave as what economists

call "price takers", essentially treating their environment as fixed, even when the collective behavior that this produces plays a large role in shaping the environment.) Early evidence suggests that in these settings many important aspects of human behavior can be captured with surprisingly simple foresight-free learning models. While these foresight-free models may shed little insight into human *thinking*, we believe that they shed light on important behavioral questions, about how people learn to respond to the incentives in their environment. They capture nontrivial properties of human behavior.

References

[1] J. Andreoni, J.H. Miller, Rational cooperation in the finitely repeated Prisoner's dilemma: Experimental evidence, The Economic Journal 103 (1993) 570–585.
[2] D. Ariely, A. Ockenfels, A.E. Roth, An experimental analysis of ending rules in Internet auctions, RAND Journal of Economics 36 (4) (2005) 891–908.
[3] G. Barron, I. Erev, Small feedback based decisions and their limited correspondence to description based decisions, Journal of Behavioral Decision Making 16 (2003) 215–233.
[4] Y. Bereby-Meyer, A.E. Roth, Learning in noisy games: Partial reinforcement and the sustainability of cooperation, American Economic Review 1 96 (4) (2006) 1029–1042.
[5] C. Bolton, O. Harris, Strategic experimentation, Econometrica 67 (1999) 349–374.
[6] G. Bornstein, E. Winter, H. Goren, Experimental study of repeated team-games, European Journal of Political Economy 12 (1996) 629–639.
[7] M.W. Cripps, J.C. Ely, G.J. Mailath, L. Samuelson, Common learning, Working paper, 2005.
[8] H. Danieli, Social dilemmas and the design of simulators, MSc Thesis, Technion, 2000 (in Hebrew).
[9] J. Denrell, J.G. March, Adaptation as information restriction: The hot stove effect, Organization Science 12 (2001) 523–538.
[10] I. Erev, On the weighting of rare events and the economics of small decisions, in: S.H. Oda (Ed.), Advances in Experimental Economics, in: Lecture Notes in Economics and Mathematical Systems, vol. 590, Springer, 2007, in press.
[11] I. Erev, G. Barron, On adaptation, maximization and reinforcement learning among cognitive strategies, Psychological Review 112 (4) (2005) 912–931.
[12] I. Erev, A. Rapoport, Magic, reinforcement learning and coordination in a market entry game, Games and Economic Behavior 23 (1998) 146–175.
[13] I. Erev, A.E. Roth, Predicting how people play games: Reinforcement learning in games with unique strategy equilibrium, American Economic Review 88 (1998) 848–881.
[14] I. Erev, A.E. Roth, On simple reinforcement learning models and reciprocation in the prisoner dilemma game, in: G. Gigerenzer, R. Selten (Eds.), The Adaptive Toolbox, MIT Press, Cambridge, MA, 2001, pp. 215–232.
[15] I. Erev, A.E. Roth, R.L. Slonim, G. Barron, Learning and equilibrium as useful approximations: Accuracy of prediction on randomly selected constant sum games, Economic Theory (2007), in press.
[16] E. Ert, I. Erev, Replicated alternatives and the role of confusion, chasing, and regret in decisions from experience, Journal of Behavioral Decision Making (2007), in press.
[17] D. Fudenberg, D. Levine, Theory of Learning in Games, MIT Press, Cambridge, MA, 1998.
[18] E. Haruvy, I. Erev, D. Sonsino, The medium prizes paradox: Evidence from a simulated casino, Journal of Risk and Uncertainty 22 (2001) 251–261.
[19] D. Kahneman, A. Tversky, Prospect theory: An analysis of decision under risk, Econometrica 47 (1979) 263–291.
[20] J.H. Kagel, A.E. Roth, The dynamics of reorganization in matching markets: A laboratory experiment motivated by a natural experiment, Quarterly Journal of Economics (2000) 201–235.
[21] D. Malcolm, B. Lieberman, The behavior of responsive individuals playing a two-person, zero-sum game requiring the use of mixed strategies, Psychonomic Science 12 (1965) 373–374.
[22] J.L. Myers, E. Sadler, Effects of range of payoffs as a variable in risk taking, Journal of Experimental Psychology 60 (1960) 306–309.
[23] B. O'Neill, Nonmetric test of the minimax theory of two-person zerosum games, Proceedings of the National Academy of Sciences USA (1987) 2106–2109.
[24] A. Rapoport, A.M. Chammah, Prisoner's Dilemma: A Study in Conflict and Cooperation, University of Michigan Press, Ann Arbor, 1965.
[25] J. Robinson, An iterative method of solving a game, Annals of Mathematics 54 (1951) 296–301.
[26] A.E. Roth, I. Erev, Learning in extensive-form games: Experimental data and simple dynamic models in intermediate term, Games and Economic Behavior 8 (1995) 164–212.
[27] R. Selten, R. Stoecker, End behavior in sequences of finite prisoner's dilemma supergames: A learning theory approach, Journal of Economic Behavior and Organizations 7 (1986) 47–70.
[28] Y. Shoham, R. Powers, T. Grenager, If multi-agent learning is the answer, what is the question? Artificial Intelligence 171 (7) (2007) 365–377, this issue.
[29] P. Suppes, R.C. Atkinson, Markov Learning Models for Multiperson Interactions, Stanford University Press, 1960.