Efficiency and Traps of Social Learning: Some Computational Experiments

Yuya Sasaki*

Department of Economics, Utah State University
3530 Old Main Hill, Logan, UT 84322-3530

Abstract

Social learning [2] exhibits characteristics unobservable in heterogeneous-agent selection dynamics in terms of stochastically stable set and the speed of convergence. We study representative advantage and disadvantage of social learning, and examine the results of computational experiments in a simple environment by applying the agent-based reinforcement learning algorithm for shared knowledge. This paper sets up a starting point.

Keywords: agent-based, reinforcement learning, social learning.

1. Introduction

The social learning theory [2], as opposed to the behaviorism, assumes that an agent can learn without a change in behavior, but through observation alone in the extreme end. Social learning becomes beneficial especially when agents face with a vast state space through which to make actions. While psychological and "micro" aspects of the social learning theory (eg. [5]) are an important issue, economists may often be interested in aggregate effects particularly with respect to the speed of evolution and the choices that agents take in equilibrium.

[3] studied a social learning model in which agents choose one of two technologies, and concluded that social learning mechanism tends to enable overall superior technology to be adopted in the long run. [4] extends [3] by adding more complexity, and their "must-see" restriction imposed on agents has an adverse effect in the aggregate as less popular alternatives, which might actually yield better payoffs, can phase out.

In this paper, we set up a naïve social learning environment in which information is uniformly shared by all the agents. (For more complex information diffusion mechanism in the spatial context, see [1] for example.) With this, we briefly discuss how this information sharing improves the efficiency or the speed of social adoption of the superior strategy as studied by [3]. On the other hand, we observe that agents' adherence to the popularly superior actions leads to possible ignorance of the unexplored superior actions as in [4]. To examine the latter, we employ the ε-greedy assumption on agents' decisions. That is, an agent takes the best action according to the social knowledge with probability $1-\varepsilon$; otherwise explores.

2. The Model

2.1. Learning

We consider n strategies. Let $P = \{1, \ldots, n\}$ be the set of strategy indices and Θ^{n-1} be the n-1-dimensional simplex. Let $r : P \times \Theta^{n-1} \to R^1$ be a payoff function.

We assume that agents can obtain the information of the state in the last time step, but no information of the current state. Let $\Omega = P \times \Theta^{n-1}$ be the set of strategy-state pairs, and $\Lambda = \{(Q^t(i))_{n\times 1} \mid i \in \mathrm{P}\}$ be the set such that $Q^t(i)$ is an agent's estimate of the payoff with strategy i. With these, an agent computes the estimated payoff by a function $F : \Omega \to \Lambda$. This function needs to be modifiable so that agents can learn. We adopt a tabular state space for the sake of simplicity in implementing computational realization of this updatable function. Suppose that Θ^{n-1} is partitioned such that $\bigcup_l \Theta_l^{n-1} = \Theta^{n-1}$, and let $L = \{l \mid \Theta_l^{n-1} \subset \Theta^{n-1}\}$ be the set of discrete state indices. Define a new set $\overline{\Omega} = \{(i', l) \mid i' \in \mathrm{P}, \Theta_l^{n-1} \subset \Theta^{n-1}\}$, and with a discrete indexing by $l \in L$, the former function can be rewritten as $\overline{F} : \overline{\Omega} \to \Lambda$.

* This paper was motivated by the comments made by Robert Axtel for my presentation in CIEF 2003 and the review reports of an anonymous referee of the post-conference publication, both of whom mentioned the importance of social learning in the author's previous work of multi-agent traffic routing model.

Note that, with set $\overline{\Omega}$ instead of Ω, the function may be represented by tile coding; this is for the sake of computational simplicity. Let $Q^t(i,l)$ denote the value of $\overline{F}(i,l)$ at time *t*. In considering the strategy to be implemented in time *t*+1, an agent should ideally choose the superior strategy given by $i^* = \arg\max_{i'} Q^t(i',l)$ for a given l such that the current state (frequency vector, which we denote by $\mathbf{x}^t$) is contained in Θ_l^{n-1}. Agents occasionally need to "explore" to take one of inferior strategies $i \neq i^*$ by a probability $\varepsilon \in [0,0.5)$. Another reason for exploration is to let agents learn for most, if not all, of the strategy-state pairs, as will be discussed later. That is, exploration makes it possible for $\overline{F}$ to be modified in an online fashion not only for i^* but also for all $i \neq i^*$. In the field of computational intelligence, this type of algorithm is called the ε-greedy method.

A straightforward way to realize $\overline{F}$ is to update $Q^t(i,l)$ based on the average payoff for strategy-state pair (i,l) from all the past experiences. Since the estimate error is given by $r(i^t,\mathbf{x}^t) - Q(i^t,l^t)$ where i^t is the strategy that the agent actually took at time *t*, and l^t is the index such that $\mathbf{x}^t \in \Theta_{l^t}^{n-1}$. Hence, the update rule of $\overline{F}$ is given in the form

$$Q(i,l) := [r(i^t,\mathbf{x}^t) - Q(i^t,l^t)]/\tau_{i,l} + Q(i,l)$$

where $\tau_{i,l}$ is the number of times the update has occurred for the given strategy-state pair (i,l) including the current time. It can be easily checked that this rule is equivalent with averaging of all the past payoffs for (*i*, *l*). Notice that exploration of inferior strategies $i \neq i^*$ is crucial here, since only $Q(i^*,l)$ which by chance currently stores the greatest value would be continuously updated otherwise. [6] discusses the importance of exploration and ε by examples. By substituting $\alpha \in (0,1)$ for $1/\tau_{i,l}$ in the above update rule, we obtain a recency-weighted learning update rule

$$Q(i,l) := \alpha[r(i^t,\mathbf{x}^t) - Q(i,l)] + Q(i,l).$$

This parameter, α, is often called learning rate in computational reinforcement learning. We may also call it the degree of recency weighting.

2.2. ε-greediness and exploration

Suppose that agents perfectly share the information. For example, a tribal group may have the social system of information sharing to ensure that nobody deviates from a Pareto dominant state. Today, information technology enables active participations in information sharing. If the play is serious to them, they have an incentive to willingly obtain the shared knowledge, even if not forced so. Such situations can be realized by $\overline{F}$ and thus $Q(i,l)$ shared by all the agents in the game. While an agent updates $Q(i,l)$, another agent can use this updated information $Q(i,l)$ to estimate a payoff by $\overline{F}$.

Suppose ε is constant throughout the game. Then, we have the frequency of exploring agents given by

$$x^{<\varepsilon>} \sim \mathrm{Bin}(N,\varepsilon)/N$$

where $\mathrm{Bin}(\cdot)$ is the binomial distribution and *N* is the finite number of agents. Then, the frequency of the agents adopting the currently superior strategy i^* is given by $x_{i^*} = 1 - x^{<\varepsilon>}$. The frequencies of all the other strategies are given by uniform distribution. See Algorithm 1 in appendix for more details about the transition rule.

Thus, the frequency of agents taking the currently superior strategy can take any value in {0,1/*N*,2/N, … ,1} with strictly positive probabilities, since *N* takes a finite integer value. For a two-strategy play, for example, this implies that the values of $Q(i,l)$ for all the states $\mathbf{x} \in \Theta^{n-1}$ are updated infinitely many times in the limit as *t* approaches infinity. This argument is also true for *n*-strategy plays.

3. Socially Superior Strategies and Effects of Exploration

For an analytical purpose, we let $N \to \infty$ for the moment. The exploration model described in Sec. 2.2 gives a certain probability distribution $\Pr_{(i^*,\varepsilon)}: \Theta^{n-1} \to \mathbf{R}$ conditional on the currently superior strategy i^*, which is yet to be defined, and the exploration probability ε. The social expected payoff is given by

$$\int_{\mathbf{x} \in \Theta^{n-1}} Q(i^*, l \mid \mathbf{x} \in \Theta_l^{n-1}) \cdot \Pr_{i^*,\varepsilon}(\mathbf{x}) d\mathbf{x},$$

and thusly the currently superior strategy is

$$i^* \equiv \arg\max_i \int_{\mathbf{x} \in \Theta^{n-1}} Q(i, l \mid \mathbf{x} \in \Theta_l^{n-1}) \cdot \Pr_{i,\varepsilon}(\mathbf{x}) d\mathbf{x}.$$

The actual superior strategy, however, is given by

$$i^{**} \equiv \arg\max_i \int_{\mathbf{x} \in \Theta^{n-1}} r(i,\mathbf{x}) \cdot \Pr_{i,\varepsilon}(\mathbf{x}) d\mathbf{x}.$$

The efficiency of social learning under the ε-greedy assumption is characterized by the convergence of i^* to i^{**}.

The speed of this convergence is certainly one criteria to evaluate the efficiency of social learning. This is, in fact, obvious. Consider the case in which *Q* is shared by all the agents and $\overline{F}$ is updated by all of them in each iteration, and consider the heterogeneous-learning in which each agent possess unique *Q* updated only by herself each time. The former certainly expedites the convergence of *Q* to *r*, thus of i^* to i^{**}.

On the other hand, the social learning can semi-permanently prevent this convergence from being realized, especially when explorations are rare, ie. ε is

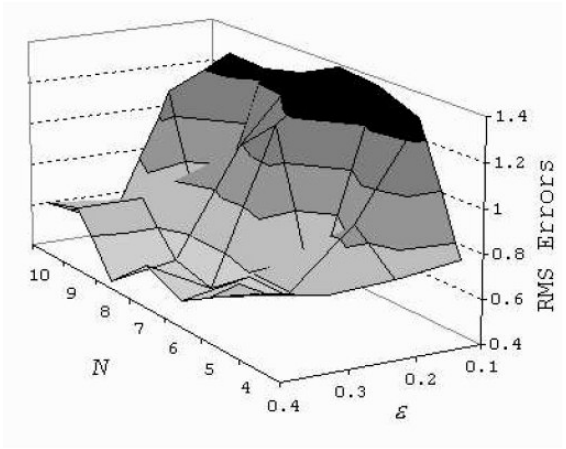

Fig. 1: RMS errors of the social knowledge Q from r, as a result of agent-based simulations.

small. Examples are perhaps more appealing to show this phenomenon.

Example 1. Consider the payoff function r is linear; the payoff matrix is given by 1 for each diagonal element and 0.5 for others. If we were dealing with a non-cooperative game, we would find two boundary equilibria and an interior equilibrium in the middle. In the ε-greedy assumption, however, we have stochastically stable points at $(1-x^{<\varepsilon>}, x^{<\varepsilon>})$ and $(x^{<\varepsilon>}, 1-x^{<\varepsilon>})$ yielding the identical social payoff. If ε is arbitrarily small, neighbors of only one of these stochastically stable points are frequently visited, and thus are learned about. In other words, social knowledge is not updated for the other half of the simplex. We ran agent-based simulations varying the number of agents from 4 to 10 and the exploration probability from 0.1 to 0.4. The results are shown in Fig. 1. The efficiency of social learning does not seem to be affected by the number of agents, but seens heavily dependent on the exploration rate.

This example shows the significance of exploration to let the social knowledge to cover the entire space. We now provide the following argument.

Proposition 1. *The greater exploration probability* $\varepsilon \in (0,0.5)$ *causes more strategy-state pairs (i, l) to be visited by the society, and thus, enables a wider variety of Q(i, l) to be updated by learning.*

Proof. Let i* be the currently superior strategy, and we have $x_{i^*} = 1 - x^{<\varepsilon>}$. By the binomial distribution, the sensitivity of the variance of exploring frequency to the rate of exploration is

$$\frac{d(\sigma^2_{x^{<\varepsilon>}})}{d\varepsilon} = 1 - 2\varepsilon > 0 .$$

The state transition is computed according to Algorithm 1 in appendix. At each iteration, the frequency of each inferior strategie $i' \in P'$ is computed as

$$x_{i'}N \sim \mathrm{Bin}(\bar{x}N, 1/(n-k))$$

with the parameters defined in Algorithm 1. Let $P'' := P^n \setminus (P' \cup \{i^*\})$ for the current iteration of the algorithm. The expected value is then

$$E(x_{i'}) = \varepsilon \textstyle\sum_{i \in P} x_i - \sum_{i'' \in P''} x_{i''} .$$

The sensitivity of variance, thus, is given as

$$\frac{d(\sigma^2_{xi'})}{d\varepsilon} = \frac{dx_{i'}}{d\varepsilon} \cdot \frac{1}{N-k}\left(1 - \frac{1}{N-k}\right)$$

$$= \frac{1}{N-k}\left(1 - \frac{1}{N-k}\right) > 0$$

as desired. □

Each point in the simplex has a locally superior strategy according to the social knowledge. To denote this, we let, for each *i,*

$$\Pi_i^n = \{\mathbf{x} \in \Theta^{n-1} \mid \mathbf{x} \in \Theta_l^{n-1} \Rightarrow Q(i,l) \geq \max_i Q(i,l)\} .$$

Example 2. We now show graphical results of computational experiments on a three-strategy case where the payoff matrix is given by 1 for each diagonal element and 0.5 for others. We have $\Pi_i^n \neq \varnothing$ for each i if Q converges to r. Otherwise, the society would have missed one of the socially superior strategies. We ran computational simulations with varying ε whose results are given by Fig. 2 and 3. We project the *two*-dmensional simplex onto the x_1-x_2 plane. Fig. 2 shows the errors of social estimates Q with respect to the actual values r on this projected plane, where the darker regions are given larger errors. Fig. 3 shows Π_i^n for each i projected on this plane. Even though strategy 1 is equally socially superior, the simulation with ε=0.1 engendered $\Pi_1^n = \varnothing$, implying that the society has missed this one of the socially superior strategies. This is typifies the trap of social learning with rare explorations.

4. Conclusion

Social learning can exhibit more complex properties than heterogeneous-agent selection dynamics. Particularly, while it can expedite learning about vast state spaces enabling faster convergence of the social action to the socially optimal strategy, a social learning with rare exploration (eg. by a strict enforcement of social norms) can prohibit the society from benefiting from the best options. We have seen some examples via computational experiments.

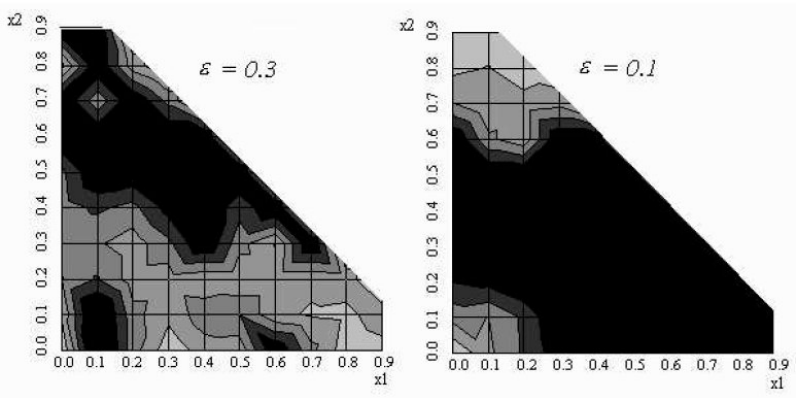

Fig. 2: Errors of social estimates Q with respect to the actual values r projected onto the x_1-x_2 plane. The darker regions indicate larger errors.

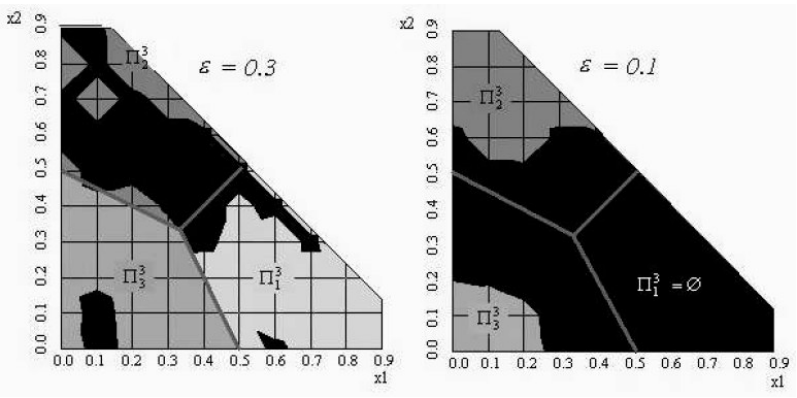

Fig. 3: The sets of locally superior points for each strategy projected onto the x_1-x_2 plane.

5. Appendix: Algorithm 1: An Estimate of State Transition

The current state $\mathbf{x}^{t+1}$ is computed by the following procedure. For $i = i^{t+1}$, $x_i = 1 - x^{<\varepsilon>}$ results where $x^{<\varepsilon>} \sim \mathrm{Bin}(N, \varepsilon)/N$. Let $\mathrm{P}' = \mathrm{P} \setminus \{i^{t+1}\}$ denote the set of currently inferior strategies. Define a variable of index k and assign the initial value k=1. Also, define a variable of frequencies computed, denoted by $\bar{x}$, and assign the initial value $\bar{x} = x^{<\varepsilon>}$. Pick one strategy $i' \in \mathrm{P}'$, and compute the frequency of agents taking this strategy by the rule

$$x_{i'} \sim \mathrm{Bin}\left(\bar{x}N, \frac{1}{n-k}\right)/N .$$

Then, increment k by one, and increment $\bar{x}$ by $x_{i'}$. Redefine the set of remaining strategies as $\mathrm{P}' := \mathrm{P}' \setminus \{i'\}$, and repeat the same process while $k < n - 1$. When $k = n - 1$, compute the frequency for one remaining strategy $i' \in \mathrm{P}'$ as $x_{i'} = 1 - \sum_{i \neq i'} x_i$.

6. References

[1] V. Bala and S. Goyal, "Learning from Neighbours," *Review of Economic Studies*, pp 595-621, 1998.

[2] A. Bandura, "Social Learning Theory," Prentice-Hall, Englewood Cliffs, NJ., 1977.

[3] G. Ellison and D. Fudenberg, "Rules of Thumb for Social Learning," *Journal of Political Economy*, pp. 612-643, 1993.

[4] G. Ellison and D. Fudenberg, "Word-of-Mouth Communication and Social Learning," *Quarterly Journal of Economics*, pp. 93-125, 1995.

[5] J.B. Rotter, "The Psychological Situation in Social Learning Theory," in D. Magnusson (Ed.), *Toward a psychology of situations: An interactional perspective.* Hillsdale, NJ: Lawrence Erlbaum, 1981.

[6] R.S. Sutton and A.G. Barto, "Reinforcement Learning: An Introduction," The MIT Press, MA, 1998.