Q learning in the minority game

M. Andrecut* and M. K. Ali
Department of Physics, University of Lethbridge, 4401 University Drive, Lethbridge, Alberta, T1K 3M4, Canada

(Received 24 July 2001; published 26 November 2001)

We present a numerical investigation of the minority game model, where the dynamics of the agents is described by the Q-learning algorithm. The numerical results show that the Q-learning dynamics is suppressing the ‘‘crowd effect,’’ which is characteristic of the minority game with standard inductive dynamics, and it converges to a stationary state that is close to the optimal Nash equilibrium of the game.

DOI: 10.1103/PhysRevE.64.067103 PACS number(s): 05.90.+m, 89.90.+n

I. INTRODUCTION

Recently, it has been shown that the minority game (MG) model can be successfully used to study the competitive interaction of complex adaptive agents in a socioeconomic environment [1,2].

In the MG model, a number of agents use a finite number of strategies to react to a finite number of public informations and interact through a collective variable whose value is fixed by all of them. The agents are choosing their strategy via a simple reinforcement learning process called inductive thinking. An essential element of this model is that the agents are rewarded if they are in minority. The goal of the game is to minimize the global loss of the agents (or to maximize the global reward). This way, the MG model can be used to obtain some qualitative understanding of more complex systems like markets. For example, a higher demand on the market will tend to increase the price and the sellers (who are the minority on the market, in this case) will be rewarded by ‘‘selling high.’’

Numerical and analytical results have shown that the MG undergoes a phase transition from an efficient phase, in which the agents coordinate their actions to minimize their loss, to an inefficient phase, where the ‘‘crowd effect’’ occurs, leading to a higher increase of the agent’s loss [3–5]. The best coordination among agents is achieved in the transition region between these two phases.

It is well known that the inductive dynamics leads to an inefficient equilibrium of the game that is far from the Nash equilibrium, i.e., the state where each agent plays the best strategy, which is minimizing the global loss and maximizing the individual utility [4,5]. In this paper we address this problem and present a numerical investigation, where the dynamics of the agents in the MG model is described by a more sophisticated learning rule, corresponding to the Q-learning algorithm.

II. THE MINORITY GAME MODEL

The MG model of the market consists of N agents that can take only two actions, such as ‘‘buy’’ and ‘‘sell’’ at each time step t. All agents have access to public information, which is an integer variable $\mu(t)$ randomly drawn at time t in the set $\{1,\ldots,P\}$. Agents have at their disposal $S\geqslant 2$ forecasting strategies, which for each value of μ suggest which action shall be taken. There are 2^P such strategies but each agent just picks S such rules randomly at the beginning of the game. These strategies (also called lookup tables) are denoted by

$$a_{n,s}^{\mu}=\pm 1, \quad n=1,\ldots,N, \quad s=1,\ldots,S, \quad \mu=1,\ldots,P. \tag{1}$$

The agents have no way of knowing what the majority will do before taking their actions (choosing their strategies).

The ‘‘payoff’’ (the gain) to each agent is given by

$$g_n(t)=-a_{n,s_n(t)}^{\mu(t)}(t)A^{\mu(t)}(t), \quad n=1,\ldots,N, \tag{2}$$

where

$$A^{\mu(t)}(t)=\sum_{n=1}^{N} a_{n,s_n(t)}^{\mu(t)}(t) \tag{3}$$

is the global variable describing the excess demand on the market at time t.

The MG interaction is described by the logical XOR function (Table I). The agents from the minority [who took the action $a(t)=-\text{sgn}(A(t))$] are rewarded with a gain $|A(t)|$, and those from majority [who took the action $a(t)=\text{sgn}(A(t))$] are punished by a loss $-|A(t)|$.

In order to choose the best strategy, each agent updates the cumulated ‘‘virtual payoff’’ for each strategy

$$U_{n,s}(t+1)=U_{n,s}(t)-a_{n,s}^{\mu(t)}(t)A^{\mu(t)}(t), \quad n=1,\ldots,N, \quad s=1,\ldots,S. \tag{4}$$

Here, ‘‘virtual payoff’’ means that this is the payoff that the agent would have received playing strategy s.

TABLE I. The MG interaction.

$\text{sgn}(a(t))$	$\text{sgn}(A(t))$	$\text{sgn}(g(t))$
-	-	-
-	+	+
+	-	+
+	+	-

*Email address: mircea.andrecut@uleth.ca

Inductive dynamics consists of assuming that agents follow that strategy with the current highest score (ties are broken by coin tossing)

$$s_n(t)=\arg\max_{s\in\{1,\dots,S\}} U_{n,s}(t), \qquad n=1,\dots,N. \tag{5}$$

The main quantity of interest is

$$\sigma^2=\langle A^2\rangle=\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}A^2(t)=-\sum_{n=1}^{N}\langle g_n\rangle, \tag{6}$$

which quantifies the fluctuations of the market and also equals the total losses of agents. Here, $A^{\mu(t)}(t)$ only depends on what agents do, so strictly speaking $\mu(t)$ has no direct impact on the market and one can replace $A^{\mu(t)}(t)$ with $A(t)$.

By symmetry we have $\langle A\rangle=0$, but it may happen that for a particular μ, the global quantity $A^{\mu(t)}(t)$ is nonzero in average, i.e., $\langle A^{\mu(t)}(t)\rangle\neq 0$. As a measure of this "asymmetry" (or "available information") one can use the quantity

$$H=\overline{\langle A^2\rangle}=\frac{1}{P}\sum_{\mu=1}^{P}\langle A^{\mu(t)}(t)\rangle^2, \tag{7}$$

such that $H=0$ when all averages vanish, $\langle A^{\mu(t)}(t)\rangle=0$.

Using the "replica method," it has been shown that, under the above mentioned inductive dynamics, H plays the role of a spin glass Hamiltonian [5]. Therefore, the ground state properties of the Hamiltonian H gives all the information on the stationary state of the system.

Numerical and analytical results have revealed the presence of a phase transition (Fig. 1) with symmetry breaking at

$$\alpha_c=\alpha_c(S)\cong S/2-0.6626\cdots, \tag{8}$$

where $\alpha=P/N$ is the main free parameter of the MG model [3–5].

If $\alpha>\alpha_c$ the system is in the asymmetric phase, $H>0$, and agents coordinate their actions to minimize their loss. The asymmetry H and the global loss σ^2 decrease with decreasing α, this means that the asymmetry in $\langle A^{\mu(t)}(t)\rangle\neq 0$ is exploited by the adaptive behavior of agents who then reduce these quantities. At $\alpha=\alpha_c$ the asymmetry vanishes and the global loss (σ^2) of the agents attains its minimum. For $\alpha<\alpha_c$ the system is in the symmetric phase, $H=0$, and the "crowd effect" occurs, leading to a higher increase of the global loss.

Another problem of interest is the understanding under what conditions adaptive learning can lead to a stationary state where each agent plays the best strategy, which is minimizing the global loss and maximizing the individual utility, i.e., the Nash equilibrium of the game [5].

For the standard MG model the stationary state is not an optimal Nash equilibrium. The asymptotic state of this dynamics is information efficient ($H=0$), but it is not optimal because when the number of agents exceeds a critical number, the market becomes symmetric and unpredictable, with large fluctuations that leads to a higher increase of the global loss ($\sigma^2\sim N$) [5].

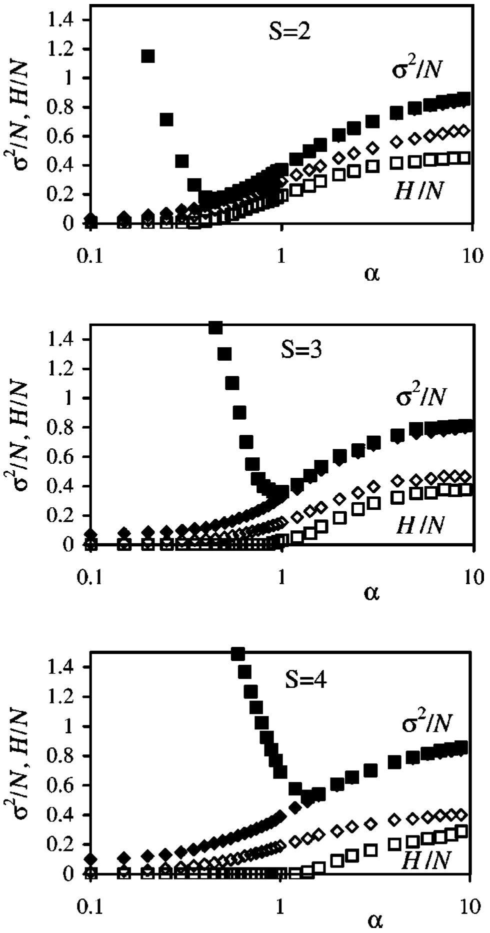

FIG. 1. Global loss σ^2/N (filled symbols) and available information H/N (open symbols) as a function of α for $S=2$, 3, and 4 from numerical simulations. The squares are for MG with standard inductive dynamics, the diamonds are for the MG with Q learning.

Using the Logit model with exponential learning and assuming that the agents account for their impact on the market, it has been shown that they attain not only an information-efficient state ($H=0$), but also an optimal Nash equilibrium with $\sigma^2\leqslant 1$ [5].

Here, we investigate a different model, where the learning rule (4) is replaced by the Q-learning algorithm [6].

III. *Q* LEARNING

Reinforcement learning (RL) is a learning technique for a class of problems in which an autonomous agent acting in a given environment improves its performance by progressively maximizing a function calculated just on the basis of a succession of scalar rewards and punishments received from the environment [7]. The agent rely only on a trial-and-error strategy and no complimentary guidance is provided for helping the exploration/exploitation of the problem space.

The agent is situated in an environment that is given as a finite Markov decision process [7]. That is, at the time t the environment is characterized by its state $s(t)$, and after each agent's action $a(t)$, the environment changes to a new state $s(t+1)$ and the agent is receiving a scalar reward $r(t+1)$.

Also, one assumes that the action and the state sets are finite. The task of the agent is to learn an optimal policy.

A policy is the mapping $\pi{:}\mathit{state}\rightarrow \mathit{action}$. An optimal policy is characterized by a maximal total discounted reward

$$R=\sum_{t=1}^{\infty} \gamma^t r(t), \quad 0<\gamma\leqslant 1. \tag{9}$$

Within the framework of dynamic programing, it has been shown that the optimal policy exists π^* and can be found iteratively [7]. If one knows the value of actions, $Q^{\pi}(s,a)$, for the present policy π, then after taking an action with the largest Q^{π}, one obtain a new policy π' with $Q^{\pi'}$ and so on. Finally, one obtains the optimal policy π^* and the corresponding values Q^*. On the other hand, if one knows all $Q^*(s,a)$, then the optimal policy is obvious given by the so-called ‘‘greedy rule’’

$$a(t)=\arg\max_{b} Q^*(s(t),b). \tag{10}$$

One of the most important breakthroughs in RL was the development of an off-policy temporal-difference algorithm known as Q learning [6,7].

The simplest form of this algorithm corresponds to the so-called one-step Q-learning equation

$$Q(s(t),a(t))\leftarrow Q(s(t),a(t))+\eta\Delta(t), \tag{11}$$

where

$$\Delta(t)=r(t+1)+\gamma\max_{b} Q(s(t+1),b)-Q(s(t),a(t)), \tag{12}$$

and η is the learning rate.

In this case, the learned action-value function Q directly approximates Q^*, the optimal action-value function, independent of the policy being followed. The only requirement is that the current policy must give a possibility to make the estimates of Q for all (s,a) pairs. This may be achieved by allowing nonoptimal actions with small probabilities. The action with the largest Q is chosen with the probability $1-\varepsilon$, and any other action is chosen with the probability $\varepsilon/(k-1)$, where k is the total number of actions in the state s (this procedure is called ε-greedy rule).

Let us see how we can implement the Q learning in the MG. First we observe that the strategies $s\in\{1,\ldots,S\}$ correspond to the states of the agent. Also, in the MG context an action means to choose a strategy $s\in\{1,\ldots,S\}$ and each state (strategy) is accessible from a given state (strategy). With this observation we can define the action-value function using three indices: $Q_{n,s,p}$, $n=1,\ldots,N$, $s=1,\ldots,S$, $p=1,\ldots,S$. The first index corresponds to the agent's number (n), the other two (s,p) are showing which states (strategies) are involved in the decision taken at time t. If $p=s$ then the agent has decided to stick with the same strategy, otherwise $(p\neq s)$ the agent has decided to change the strategy.

We would like to underline that the agents are playing ‘‘blind’’ following the ‘‘ε-greedy rule,’’ searching for their best rewarding strategy in the long run. This means that if an agent is in the state (strategy) s then at the next step the agent will choose the strategy (state) p with the probability $1-\varepsilon$ such that

$$p=\arg\max_{q} Q_{n,s,q}, \tag{13}$$

otherwise the agent will choose any strategy $Q_{n,s,q}$, $q\neq p$, with a probability $\varepsilon/(S-1)$.

One can see that at this level, the public information μ has no direct impact on the game and it is used only at the learning level. The update learning rule for the one-step Q-learning equation is given by

$$\begin{aligned} Q_{n,s,p}\leftarrow Q_{n,s,p}+\alpha[&-a^{\mu}_{n,p}(t)A^{\mu}(t)\\ &+\gamma\max_{w} Q_{n,p,w}-Q_{n,s,p}], \end{aligned} \tag{14}$$

where the reward r has been replaced with the ‘‘payoff’’ to each agent (2).

The numerical simulation results are given in Fig. 1. The best results have been obtained using the following parameters: $P=128$, $\eta=1/P$, and $\gamma=0.5$. In order to obtain good statistical results the data have been averaged over 200 ‘‘measurements.’’ The number of iteration steps in each measurement was $T=10^5$ and the ε-greedy variable was decreased as $\varepsilon=1/t$, $t=1,\ldots,T$.

The Q-learning dynamics leads to a symmetry broken phase for any $\alpha>0$: $H\rightarrow 0$ when $\alpha\rightarrow 0$. Also, the fluctuations decrease, the system approaching a stationary state. However, for $N\rightarrow\infty$ $(\alpha\rightarrow 0)$ the agents are not able to disentangle completely their actions and this stationary state, characterized by $1<\sigma^2\ll N$, is not an optimal Nash equilibrium (where $\sigma^2\leqslant 1$) [5].

IV. CONCLUSIONS

We have investigated numerically the role of reinforcement learning in the minority game problem. We have shown that the poor performance of the agents (due to the occurrence of ‘‘the crowd effect’’) is almost completely suppressed if the standard inductive learning dynamics is replaced by a simple one-step Q-learning algorithm. We conclude that Q-learning agents behave almost optimally, in the sense that for $N\rightarrow\infty$ $(\alpha\rightarrow 0)$ they converge to a stationary state with a nonvanishing σ^2/N, which is close to the optimal Nash equilibrium.

ACKNOWLEDGMENTS

This work has been supported through a grant to M.K. Ali from the Defense Research Establishment Suffield. The authors would also like to thank to Dr. S. Barton for his continuing support and helpful discussions.

[1] D. Challet and Y.-C. Zhang, Physica A **246**, 407 (1997).
[2] D. Challet and Y.-C. Zhang, Physica A **256**, 514 (1998).
[3] R. Savit, R. Manuca, and R. Riolo, Phys. Rev. Lett. **82**, 2203 (1999).
[4] D. Challet and M. Marsili, Phys. Rev. E **60**, R6271 (1999).
[5] M. Marsili, D. Challet, and R. Zecchina, Physica A **280**, 522 (2000).
[6] C. J. Watkins and P. Dayan, Mach. Learn. **8**, 279 (1992).
[7] R. S. Sutton and A. G. Barto, *Reinforcement Learning* (MIT Press, Cambridge, MA, 1999).