J Ethol (2001) 19:1–8

ARTICLE

Joe Yuichiro Wakano · Norio Yamamura

A simple learning strategy that realizes robust cooperation better than Pavlov in Iterated Prisoners' Dilemma

Received: June 23, 2000 / Accepted: October 1, 2000

Abstract Pavlov was proposed as a leading strategy for realizing cooperation because it dominates over a long period in evolutionary computer simulations of the Iterated Prisoners' Dilemma. However, our numerical calculations reveal that Pavlov and also any other cooperative strategy are not evolutionarily stable among all stochastic strategies with memory of only one previous move. We propose simple learning based on reinforcement. The learner changes its internal state, depending on an evaluation of whether the score in the previous round is larger than a critical value (aspiration level), which is genetically fixed. The current internal state decides the learner's move, but we found that the aspiration level determines its final behavior. The cooperative variant, having an intermediate aspiration level, is not an evolutionarily stable strategy (ESS) when evaluation is binary (good or bad). However, when the evaluation is quantified some cooperative variants can invade not only All-C, Tit-For-Tat (TFT), and Pavlov but also noncooperative variants with different aspiration levels. Moreover, they establish robust cooperation, which is evolutionarily stable against invasion by All-C, All-D, TFT, Pavlov, and noncooperative variants, and they receive a high score even when the error rate is high. Our results suggest that mutual cooperation can be maintained when players have a primitive learning ability.

Key words Prisoners' dilemma · Evolution of cooperation · Learning · Pavlov · Mathematical model · Computer simulation

J.Y. Wakano (✉) · N. Yamamura
Center for Ecological Research, Kyoto University, Kamitanakami Hiranocho, Otsu, Shiga 520-2113, Japan
Tel. +81-77-549-8200; Fax +81-77-549-8201
e-mail: joe@ecology.kyoto-u.ac.jp

Introduction

The Iterated Prisoner's Dilemma has played an important role in the study of evolution of cooperative behavior in populations of selfish agents (Axelrod and Hamilton 1981). Two players engaged in the Prisoner's Dilemma must choose between cooperation (C) and defection (D). In any round, the two players receive R points (the reward for cooperation) if both cooperate and only P points (the punishment for mutual defection) if both defect. A defector exploiting a cooperator gets T points (the temptation to defect) while the cooperator receives S points (the sucker's payoff). When $T > R > P > S$ and $2R > T + S$, it is always best to defect in a single round, and hence mutual defection is a logical result although both players receive more points in mutual cooperation. The best-studied set of score values are $T = 5$, $R = 3$, $P = 1$, $S = 0$ (e.g., Axelrod 1984; Sandholm and Crites 1995; Brauchli et al. 1999; Posch 1999), which is adopted also in this study.

When the Prisoner's Dilemma game is repeated (Iterated Prisoner's Dilemma, IPD), the most adaptive strategy is not clear. In a series of computer tournaments (Axelrod 1984), a simple strategy Tit-For-Tat (TFT) did outstandingly well. TFT plays C in the first round and then plays its opponent's previous move. Nowak and Sigmund (1993a) ran an integrated evolutionary simulation including error (which was not considered in Axelrod's tournaments) among stochastic strategies whose moves were dependent only on the moves of the previous round. Such strategies can be written as (p_1,p_2,p_3,p_4), each component of which corresponds to the probability of playing C when the previous round is CC, CD, DC, or DD, respectively (the former letter represents its own move while the latter is the opponent's move). The winning strategy was Pavlov, which plays C only after mutual cooperation or mutual defection. Significantly, Pavlov realizes mutual cooperation when matched against another Pavlov. However, Pavlov is invaded by All-D and hence is not an evolutionarily stable strategy (ESS) (Nowak and Sigmund 1993a; Stephens et al. 1995). We study here whether an evolutionarily stable

cooperative strategy exists among (p_1,p_2,p_3,p_4) type strategies. We also examine the behavior of Pavlov under high error rates because Nowak and Sigmund (1993a) conducted simulation only under 1% error rate.

Pavlov is referred to as "win–stay lose–shift" strategy; this is because Pavlov (1,0,0,1) can be interpreted as "keep the same play when rewarded (T, R) but change play when punished (P, S)." This response is an intuitively understandable standard of action. Generally, the learning rule of "positive learning: repeat when the result in the previous round is good, and negative learning: change when the result in the previous round is bad" is called reinforcement learning, which is widely observed in many kinds of creatures including apes, mice, rats, pigeons (Papini 1997), toads (Brattstrom 1990), and spiders (Whitehouse 1997).

Recently, there have been some studies in which the reinforcement learning rule is adopted in IPD. Stephens and Clements (1998) surveyed the condition under which cooperation is achieved as a result of learning. Posch (1999) performed evolutionary simulation among reinforcement learning strategies and found that cooperation is established. Sandholm and Crites (1995) introduced Q-learning as a player's learning method. However, these studies lack evolutionary stability analysis. Especially, it is interesting to analyze evolutionary stability against invasion of the well-known simple classic strategy, TFT, which cannot be represented in the form of a reinforcement learning rule. Evolationary stability analysis may also reveal to what extent learning strategy is stable against invasion of unconditional strategies (All-C and All-D). If unconditional strategy scores almost the same as learning strategy, a learning strategy cannot evolve when cost of learning is introduced. In this article, we propose simple reinforcement learning strategies and show that some of them are a cooperative learning strategy that establishes more robust reciprocal cooperation than Pavlov does. We perform evolutionarily stability analysis to study whether such cooperative learning strategy is stable against invasion of non-cooperative learning strategies, classic strategies (TFT and Pavlov) and unconditional strategies (All-C and All-D).

Model and results

Definition of evolutionarily stable strategy (ESS)

To perform evolutionarily stability analysis, we must define ESS. In this article, we use Maynard-Smith and Price's (1973) definition. There is no ESS in Iterated Prisoners' Dilemma when all possible strategies are considered (Boyd and Lorberbaum 1987; Lorberbaum 1994). However, when the strategy set is restricted, ESS may exist. If a strategy is stable against invasion of any other strategy, it is called as an ESS. Thus, strategy $x^* \in X$ is an ESS, if and only if

$$E(x,x^*) \leq E(x^*,x^*) \tag{1}$$

$$E(x^*,x) > E(x,x) \quad \textit{if } E(x,x^*) = E(x^*,x^*) \tag{2}$$

for any strategy x belonging to the strategy set X. $E(x,y)$ is the payoff of strategy x when matched against strategy y. When we refer to ESS, it is very important to show the strategy set. We use term "ESS within set X" in this article.

Performance of classic strategies

The average points of a (p_1,p_2,p_3,p_4) type strategy against the same strategy in an infinitely repeated game can be analytically calculated (Nowak and Sigmund 1993b). Using the same method with changing error rates, we calculated the average scores of populations in which individuals take the same strategy, All-C, Pavlov, TFT, All-D, or GTFT (see Table 1 for their definition and scores).

When the game is repeated long enough, the existence of errors makes the average score of a TFT population much smaller than R, the point for mutual cooperation. This change occurs because just one error pushes them out from the initial CC state and they play CD and DC repeatedly until they return to the CC state again with another error. On the other hand, Pavlov can soon come back to CC through DD when CD or DC is played due to an error. In this sense, the CC state is stable in Pavlov population, and this is the reason why reciprocal cooperation is finally achieved by Pavlov in the evolutionary simulation (Nowak and Sigmund 1993a).

The average score of a Pavlov population is, however, lower than that of an All-C population and the difference increases as the error rate increases. To enjoy the maximum benefit of cooperation when the error rate is high, the player must play C after CD, which is caused by the opponent's error. Generous Tit-For-Tat (GTFT) plays C after CD at a certain probability and hence receives a higher score than Pavlov. However, a strategy that plays C after CD is completely exploited by All-D and is not evolutionarily stable.

Let us consider a set of deterministic strategies that have the memory of only one previous move (2^4 possibilities, C or D, after CC, CD, DC, and DD). There are only two ESSs within this set, All-D and GRIM = (1,0,0,0) (Nowak and Sigmund 1993b). GRIM is a grim strategy in the sense that it never forgives the opponent's defection. Under the presence of the error, the stable state of the GRIM population as well as the All-D population is mutual defection. We have derived ESSs within (p_1,p_2,p_3,p_4) type stochastic strategies by the following method. We considered 41^4

Table 1. Average scores of populations for infinitely iterated Prisoners' Dilemma

Error rate (%)	All-C (1,1,1,1)	Pavlov (1,0,0,1)	TFT (1,0,1,0)	All-D (0,0,0,0)	GTFT (1,1/3,1,1/3)
0	3.000	3.000	3.000	1.000	3.000
1	2.990	2.951	2.010	1.030	2.958
5	2.948	2.777	2.048	1.148	2.820
10	2.890	2.602	2.090	1.290	2.692
20	2.760	2.376	2.160	1.560	2.520

All-C, always cooperate; TFT, Tit-For-Tat; All-D, always defect; GTFT, generous TFT

strategies, p_is of each strategy taking 41 discrete values spaced evenly between 0 and 1. First, results of all possible matches are numerically calculated, that is, $E(s_1,s_2)$ for all combinations of s_1 and s_2 is calculated and stored in memory. Second, for each strategy s*, we investigated whether $E(s,s^*)$, $E(s^*,s^*)$, and (if necessary) $E(s^*,s)$ meet the ESS definition or not. Results of our numerical calculation suggested that only All-D and GRIM are ESSs. GTFT and Pavlov, which are known to realize cooperation, were not ESSs within this set. Nowak and Sigmund claimed that a Pavlov-like strategy (0.999, 0.001, 0.001, 0.995) could not be invaded by All-D (Nowak and Sigmund 1993a), but we confirmed that the Pavlov-like strategy is invaded by Pavlov, which is invaded by All-D. In conclusion, within the most frequently studied strategy set (deterministic and stochastic strategies with one-move memory), no strategy that realizes cooperation is an ESS.

Introducing learning players

Pavlov does not explain actual cooperative behaviors for two reasons: it is not an ESS and it does not behave very cooperatively under high error rates (see Table 1). Therefore, we propose a simple learning player that may explain cooperative behaviors better than Pavlov. We suppose that learning players make a decision based on their internal states. The internal state continues to change gradually according to past experience (results of games), and the accumulation of small changes leads to a switch of action when the internal state crosses the threshold level (=0 in our computer simulation). So, the internal state continues to change even when the same result of the game is repeated. We denote the internal state by h. The learning player plays C (cooperation) when $h \geq 0$, otherwise D (defection). We do not assume an upper or lower limit in internal state h in the computer simulation; the large absolute values represent deep faith in cooperation or defection.

When we model the rule of reinforcement learning, it is important how many points are necessary for a player to judge whether the result is good. We assume that the threshold score is genetically fixed as aspiration level s. As the dynamics of s proceed on the evolutionary time scale, s is set to a constant value when the dynamics of h (learning process) are investigated. The dynamics of h are based on the rule of reinforcement learning and defined as follows. If C is played and the resulting score f is larger than s, cooperation is affirmatively learned. That is, the player increases h to become more cooperative. If C is played and the resulting score f is smaller than s, cooperation is negatively learned. That is, the player decreases h to become less cooperative. If D is played and the resulting score f is larger than s, defection is affirmatively learned, leading to decrease in h, and vice versa.

Digital learner

First, we suppose the most simple case in which the change in h per round is a constant value a. We name this digital learner, or DL. The formulation of its learning method is as follows:

$$\Delta h = a \cdot \mathrm{sgn}(f - s) \cdot \mathrm{sgn}(h)$$

$$\mathrm{sgn}(x) \equiv \begin{cases} 1 (x \geq 0) \\ -1 (x < 0) \end{cases}$$

As a gives only the scale of the player's internal variable h, the behavior of the player is independent of a. Aspiration level s does not influence the learning process as long as the sign of $f - s$ is the same. The only thing that matters is whether s is larger than each of T, R, P, and S. Thus, there are only five independent learners ($s \leq 0$, $0 < s \leq 1$, $1 < s \leq 3$, $3 < s \leq 5$, $s > 5$). They are abbreviated as DL($s \leq 0$), DL($0 < s \leq 1$), and so on. When $1 < s \leq 3$, the behavior is similar to Pavlov as the player is satisfied with T and R while not satisfied with P and S. However, DL($1 < s \leq 3$) is still different from Pavlov because it plays based on its internal state h.

An error occurs with probability e when the player takes the opposite action from that which is expected from the h value. Learning is done according to the foregoing expression even when an error occurs, thus assuming the player does not notice their error.

The sign of the initial h value determines the initial action. A major reason why many strategies that realize reciprocal cooperation are not evolutionarily stable is that they cannot prevent invasion of All-C by genetic drift. If the initial action is D, then the player may be able to learn to exploit All-C. We assume that the initial h value is $-a/2$ and h is reset to the initial value every time the opponent changes. With this initial state (a naive defective state), the player may change the next action to C, reacting to the initial move of the opponent. The initial action of Pavlov is also assumed to be D.

When the number of repetitions is small, the result is nearly random because the match ends before sufficient learning is done. We repeat games 10000 times because preliminary tests showed that such a number of times is necessary to clarify the effects of learning. For example, the action of DL($1 < s \leq 3$) is shown in Table 2. This player acts like Pavlov but the cooperation realized among DLs($1 < s \leq 3$) is more stable. It is so because the h values of both players become very large when cooperation continues for a long time and hence they cooperate, even though a player's h value decreases a little, after its opponent makes an occasional error to play D (Pavlov comes back to CC through DD as CC, DC, DD, CC, . . .). As shown in Table 1, the average score for populations of Pavlov or GTFT decreases as the error rate increases: 2.951 (1% error) to 2.602 (10% error) for Pavlov and 2.958 (1% error) to 2.692 (10% error) for GTFT. On the other hand, the average score in population of DL($1 < s \leq 3$) decreases little as the error rate increases: 2.985 (1% error) to 2.865 (10% error) (see Table 3).

Which strategy is an ESS within the set of five independent digital learners? This can be determined by examining scores of each player when it plays IPD against different

Table 2. Comparison of digital learner ($1 < s \leq 3$, $a = 2$) and Pavlov

Strategy	Opponent	Actual realization of game (h)	Average payoff
DL($1 < s \leq 3$)	All-D	DD(-1), CD(1), DD(-1), CD(1), DD(-1), ...	0.5
	All-C	DC(-1), DC(-3), DC(-5), DC(-7), DC(-9), ...	5
	Pavlov	DD(-1), CC(1), CC(3), CC(5), CC(7), ...	3
	TFT	DC(-1), DD(-3), DD(-1), CD(1), DC(-1), ...	1.75
	DL($1 < s \leq 3$)	DD(-1), CC(1), CC(3), CC(5), CC(7), ...	3
Pavlov	All-D	DD, CD, DD, CD, DD, ...	0.5
	All-C	DC, DC, DC, DC, DC, ...	5
	Pavlov	DD, CC, CC, CC, CC, ...	3
	TFT	DC, DD, CD, DC, DD, ...	2
	DL($1 < s \leq 3$)	DD, CC, CC, CC, CC, ...	3

DL, digital learner

Table 3. Results of matches among digital learners

	Player (mutant)	$s \leq 0$	$0 < s \leq 1$	Opponent $1 < s \leq 3$	(Wild type) $3 < s \leq 5$	$s > 5$
(a) 1% error	$s \leq 0$	1.030049	1.030053	2.965461	2.965461	2.985005
	$0 < s \leq 1$	1.030048	1.030052	2.985288	2.965454	2.974166
	$1 < s \leq 3$	0.535146	0.535152	2.984959	0.584596	1.519351
	$3 < s \leq 5$	0.539876	0.53988	2.943325	1.738342	2.502725
	$s > 5$	0.535053	0.545318	3.954128	1.258617	2.010029
(b) 10% error	$s \leq 0$	1.290936	1.290994	2.851567	2.709865	2.849876
	$0 < s \leq 1$	1.290922	1.290977	2.851499	2.709804	2.788106
	$1 < s \leq 3$	0.848479	0.848549	2.865511	1.243602	1.848385
	$3 < s \leq 5$	0.888843	0.88891	2.482726	1.910963	2.405803
	$s > 5$	0.850024	0.98887	3.526731	1.601163	2.090827

and the same strategies (note that the initial value of h is reset when the opponent changes). Matches among five different DLs were conducted by computer simulation and the results are shown in Table 3. When the same strategies are matched, mutual cooperation is realized only among DL($1 < s \leq 3$) because the score is near to 3 points. However, when the wild type in a population is DL($1 < s \leq 3$), the mutant DL($s > 5$) receives more points than the wild type for both high and low error rate cases. Therefore, DL($1 < s \leq 3$) is not an ESS. On the other hand, when DL($s \leq 0$) is the wild type, the score of DL($s \leq 0$) is always higher than the score of any mutant. Table 3 shows that DL($s \leq 0$) is the only ESS for both high and low error rates. DL($s \leq 0$) is the strategy that plays D in every round because it judges any result as "good" and the initial h value is negative. We do not present here the results of IPD between DL and the classic strategies because the evolutionarily stable DL does not show cooperative behavior.

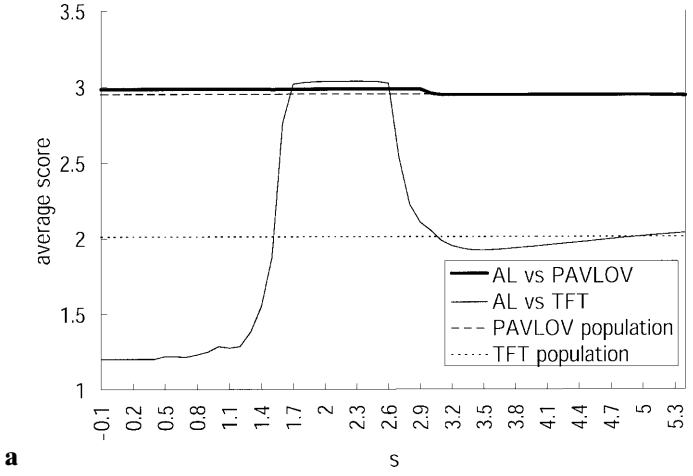
Analog learner

The way of learning for the digital learner is binary, good or bad. So, the impact on learning process is the same whether DL($1 < s \leq 3$) gets 5 points or 3 points. However, the magnitude of satisfaction would normally be different for different payoffs. Hence, let us consider another model in which the impact is proportional to the difference between resulting score f and aspiration level s:

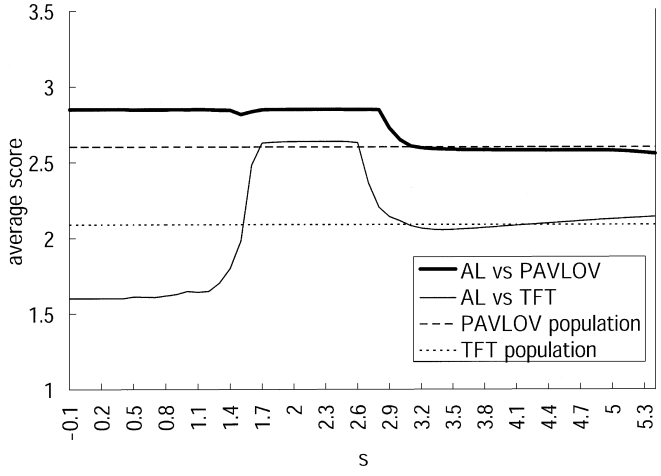
$$\Delta h = a(f - s)\cdot \mathrm{sgn}(h)$$

We refer to this learning process as the analog learner (AL) model. In this model, a small difference in the genetic parameter s produces different learning processes. In the computer simulation, we used 56 s values from -0.1 to 5.4 at 0.1 intervals. Assumptions for the initial h value, the error rate e, the learning method in the case of error, and the number of games are the same as for the DL model. Obviously, the behavior of AL is also independent of the a value.

First, we survey whether AL can invade classic strategies. It is clear that AL cannot invade All-D because no strategy can receive more points than All-D when matched against All-D. ALs for $s < 5$ invade All-C and are stable against All-C. The reason is as follows. When ALs for $s < 5$ are matched against All-C, they play D in the first round, and the result of the first round is DC. As ALs for $s < 5$ judge DC as good, they continue to learn to play D with decreasing h and hence the exploitation never ceases. We calculated the average scores of ALs with different s values when they are matched against TFT or Pavlov. The average scores for different s values are shown with the average score that TFT or Pavlov gets against itself (Fig. 1). When an error rate is 1% (Fig. 1a), ALs for $s \leq 3.0$ make a slightly better score against Pavlov than the average score of a Pavlov population and hence ALs for $s \leq 3.0$ can invade a Pavlov population. ALs for $1.6 \leq s \leq 3.0$ make a better score against TFT than the average score of a TFT population and hence ALs for $1.6 \leq s \leq 3.0$ can invade a TFT

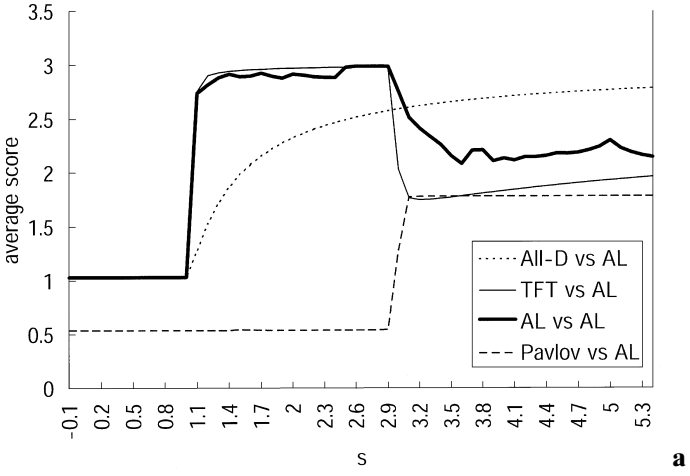


a

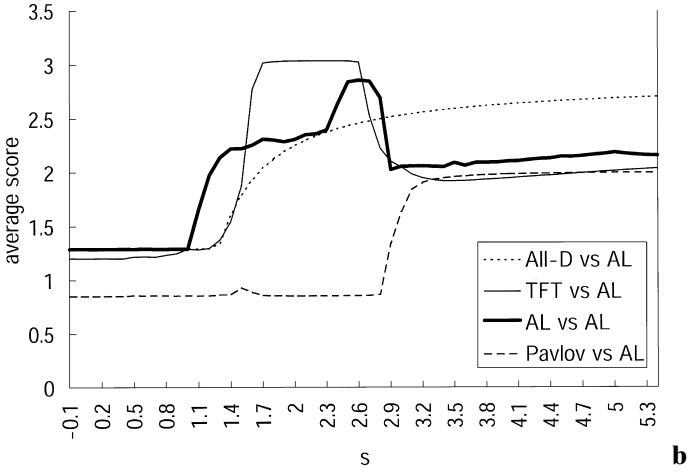


b

Fig. 1a,b. Invasion of analog learner (AL). The average scores of ALs with different s values against Pavlov and Tit-For-Tat (TFT) are shown with the average score of a Pavlov and a TFT population. AL can invade Pavlov or TFT when AL receives more points than the average score of a Pavlov or a TFT population when matched against Pavlov or TFT. The average value of 10000 runs with different random number seeds is shown. Error rate: (**a**) 1%; (**b**) 10%



a



b

Fig. 2a,b. Stability of analog learner (AL). The average scores of All-D, Pavlov, and TFT against ALs with different s values are shown with the average score of the AL population. All-D, Pavlov, or TFT can invade AL when it receives more points than the average score of the AL population. The average value of 10000 runs with different random number seeds is shown. Error rate: (**a**) 1%; (**b**) 10%

population. When an error rate is 10% (Fig. 1b), ALs for $s \leq 3.1$ can invade a Pavlov population and ALs for $1.6 \leq s \leq 3.0$ can invade a TFT population. In conclusion, analog learners for $1.6 \leq s \leq 3.0$ can invade All-C, TFT, and Pavlov in both high and low error rate situations.

Second, to study the stability of a AL population, we calculated the average score which All-D, TFT, or Pavlov gets against ALs with different s values and the average scores of the AL populations (Fig. 2). For $1.6 \leq s \leq 3$ where AL can invade All-C, TFT, and Pavlov, AL populations are stable against invasion of All-D and Pavlov for both high and low error rate situations. AL populations are stable against invasion of TFT for $s = 2.6 - 3.0$ for a 1% error rate and for $s = 2.7 - 2.8$ for a 10% error rate. In conclusion, ALs with $s = 2.7 - 2.8$ are stable against all of All-D, Pavlov, and TFT in either case of 1% or 10% error.

The average scores of AL populations for $1 < s < 3$ are generally high. Because the score of a player experiencing the repetition of DC and CD is 2.5, the average score exceeding 2.5 implies that CC is played frequently in the population. We regard such populations as cooperative. AL plays cooperatively for $1.1 \leq s \leq 3.1$ when the error rate is 1% and for $2.4 \leq s \leq 2.9$ when the error rate is 10%. Over most of these cooperative regions, the average score of the AL population is larger than that of the Pavlov population (Figs. 1, 2). The reason why cooperative action is observed for $1 < s < 3$ is as follows. Playing C after CC is affirmatively learned only when $s < 3$. The initial outcome is always DD, and ALs for $s < 1$ continue to play D. So, only ALs for $1 < s < 3$ can depart from the initial DD state and establish stable cooperation maintained by an increasing belief in C.

To study evolutionary dynamics of the aspiration level s, the results of games among ALs with different s values were calculated and are summarized in Fig. 3. When the error rate is 1%, ALs with s values around $s = 1.2$ (s = 1.1,1.2,1.4,1.7) and around $s = 2.8$ (s = 2.5,2.6,2.7,2.8,3.0) are global ESSs (stable against any invader). The AL(s = 2.8) that makes the best average score of population is an

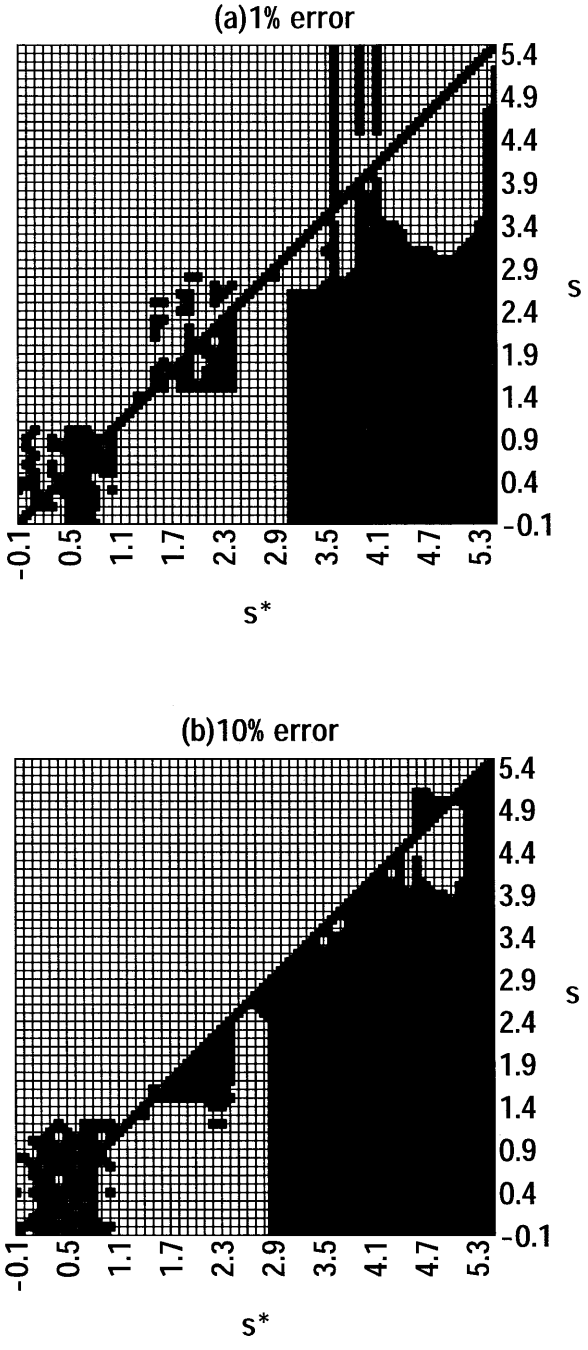


Fig. 3a,b. Results of games among ALs with different s values were calculated and summarized in a pairwise invasibility plot. Mutant AL(s = s) can invade wild type AL(s = s*) in the *gray* regions while it cannot in *white* regions. When the average score of a mutant is larger than or equal to the average score of a wild type, the invasion is assumed to be successful and so the diagonal line is always *gray*. The scores were calculated as the average of 1000 runs with different random number seeds. Error rate: (**a**) 1%; (**b**) 10%

ESS. ALs for $s \geq 3.1$ except for s = 3.6,3.9 are local ESSs (stable against neighboring invaders) but they are invaded by ALs with smaller s values. When an error rate is 10%, AL(s = 1.1,1.2,1.3) and AL(s = 2.5,2.6) are global ESSs. The AL(s = 2.6) that makes the best average score of a population is an ESS. ALs(s = 4.5,5) are local ESSs, but they are invaded by ALs($s < 3.8$). For both 1% and 10% error rates, global ESSs are always high-average scoring variants ($1 < s < 3$).

ALs($1 \leq s \leq 3$) can easily invade a TFT population (Fig. 1), but they are invaded by TFT except for a narrow region around s = 2.7 (Fig. 2). The match between ALs($1 \leq s \leq 3$) and TFT become basically mutual cooperation because only CC is a stable equilibrium. When cooperation is repeated for a long time, the h value of AL becomes very large and does not respond to D, which is played by TFT by error so it continues to play C. On the other hand, when AL plays D by error, TFT revenges by playing D once and then they come back to reciprocal cooperation. This may be the reason that TFT has a slight advantage when matched against most ALs.

When the error rate is 1%, an AL population is stable against invasion of TFT at s = 2.6,2.7,2.8,2.9 (Fig. 2a) and among these, ALs(s = 2.6,2.7,2.8) are the ESS in the evolutionary dynamics of s (Fig. 3a). At these s values, neither classical strategies (All-C, All-D, or Pavlov) can invade AL. This result suggests that ALs(s = 2.6,2.7,2.8) would finally become a majority in a population.

When the error rate is 10%, an AL population is stable against invasion of TFT at s = 2.7,2.8 (Fig. 2b). On the other hand, evolutionarily stable s values are s = 2.5,2.6 (Fig. 1b). Hence, larger s values (2.7 or 2.8) are preferred when matched against TFT while smaller s values (2.5 or 2.6) are preferred when matched against another AL. TFT cannot become a majority in a population because TFT scores very low against itself and is easily invaded by AL($1 \leq s \leq 3$). As a result of these evolutionary dynamics in case of a 10% error rate, AL(s = 2.6,2.7) may become a majority in a population with TFT at a very low frequency.

Discussion

Character of digital learner

The DL model is very similar to the model of Stephens and Clements (1998). In their model, the probability of playing C is directly changed by learning while here the internal state h is changed. They obtained the same result as for our model that there are five possible aspiration levels and cooperation is learned when only P and S are interpreted as punishment ($1 < s \leq 3$) (see Table 3). However, they did not refer to evolution of aspiration levels at all whereas we studied evolutionary dynamics among the five variants of DL. We found that the cooperative variant ($1 < s \leq 3$) is not an ESS and that there is only one ESS which is, DL($s < 0$). DL($s < 0$) acts completely like All-D. Considering that the All-D strategy is much more simple and may cost less, digital learners would be replaced by All-D and disappear soon even if it appeared by mutation.

In both DL and AL models, we assumed that there is no upper or lower limit in internal state h. If a range of h is limited, different results may occur. For example, if the h value can take only $-a/2$ or $a/2$, DL($1 < s \leq 3$) is equivalent

to Pavlov. If the range of h is small, learners cannot establish robust cooperation based on strong belief and so they cannot behave very cooperatively under high error rates. The large range of h is a key to robust cooperation.

Character of analog learner

Nowak and Sigmund showed that Pavlov continues to maintain a majority in a population for a very long period with an integrated evolutionary simulation (Nowak and Sigmund 1993a). One purpose of our study was to survey the possibility of a learning system that would replace the unstable cooperative state realized by Pavlov, which is not an ESS within the set, and attain a more solid state of mutual cooperation. Another purpose was to survey for a learning system that ameliorates the inevitable decrease in average score brought by increasing error rates. Such a learning system must be stable against invasion of more simple strategies without a long-term memory. We found that AL($s \approx 2.7$) is the one that most nearly satisfies these conditions. It is an ESS within the set of analog learners, establishes stable cooperation even when an error rate is high, and is stable against invasion of well-studied classic strategies (All-C, All-D, TFT, and Pavlov).

A major reason why many strategies that realize reciprocal cooperation are not evolutionarily stable is that they cannot prevent invasion of All-C by genetic drift. We defined the analog learner's initial move as D, so it learned to exploit All-C. Even under this restriction of initial move, we observed that learners can establish mutual cooperation. Such cooperative learners also learned to avoid being exploited by All-D. When All-D invades AL's population ($s \approx 2.7$), it scores about 2.5 points on average, which is smaller than population average score by about 0.4 points (for the reason, see next section). Thus, this cooperative population is evolutionarily stable against invasion of unconditional strategies even if the learning brings some cost so long as the cost is smaller than 0.4 points. Pavlov is cooperative and is not invaded by All-C and thus it is regarded as the leading strategy for realizing cooperation.

When cost of learning is introduced, however, Pavlov is invaded by All-D. We have shown no cooperative strategy is an ESS within stochastic strategies with one-move memory. On the other hand, we found a cooperative analog learner that is not invaded by All-C and All-D as well as TFT and Pavlov. This property is unique in AL of our model, and has never been found in other models of learning strategies (Stephens and Clements 1998; Posch 1999; Sandholm and Crites 1995).

AL comes to play only C (or D) when it has strong belief ($|h| >> 0$); this is why it can obtain high average scores by mutual cooperation under the existence of errors. The stability of cooperative variants ($1 < s < 3$) against noncooperative ALs is because they are not satisfied with CD and immediately learn to escape from the opponent's exploitation. However, there exists a strategy that presumes on these properties of AL. The anti-AL strategy continues to play C until AL has strong belief in C and plays D to exploit AL until the belief almost disappears and then plays C again so that AL's belief in C recovers. The strategy given here is more complex than AL because it estimates not only the opponent's play but also the opponent's internal state. Whether or not such a strategy could evolve depends on costs of the estimation.

Difference between digital learner and analog learner

The cooperative variant ($1 < s \leq 3$) is not an ESS within the set of digital learners while some of cooperative variants of AL are ESSs within the set of analog learners. What makes this difference? The amount of change in h value of DL per one learning process is constant, and the player's h value takes the one of one-dimensional lattice points with coordinate $(n + 1/2)\cdot a$ ($n = 0, \pm1, \pm2, \pm3, \ldots$). For example, when DL($s \leq 0$) and DL($1 < s \leq 3$) are matched, the h value of the former monotonously decreases because it judges any result as good and repeats the initial move D. Thus, the latter cannot get $R(= 3)$ points or T$(= 5)$ points and its judgment is always bad. As a result, the h value of the latter takes $-a/2$ and $a/2$ by turns and the result of the game takes DD and DC by turns. DL($s \leq 0$) takes 3 points on average and can invade by genetic drift the population of DL($1 < s \leq 3$) in which the average score is 3. This is the same mechanism by which Pavlov is invaded by All-D. The existence of an error slightly changes the actual values, but the essential mechanism remains the same.

On the other hand, the amount of change in h value of AL is dependent on the result of the previous game. For example, when AL($s = -0.1$) and AL($s = 2$) are matched, the h value of the former also monotonously decreases but the h value of the latter takes $-0.5a$, $0.5a$, $-1.5a$, $-0.5a$, $0.5a$, $-1.5a$, and so on. The result of the game is that DD repeats two times after each single DC. As a result, AL($s = -0.1$) takes 2.3 points on average and cannot invade the population of AL($s = 2$) where players behave cooperatively and get 3 points on average. This result holds so long as $1 \leq s \leq 3$. This essential difference in the learning process caused different results between DL and AL. Quantified evaluation brings learners an efficient ability to avoid being exploited by always defecting strategies.

Characteristics of our study

There are many theoretical studies about how cooperation evolves in species under situations like the Iterated Prisoner's Dilemma. Many factors have recently been introduced to this problem, such as spatial structure (Lindgren and Nordahl 1994; Ferriere and Michod 1996; Brauchli et al. 1999), rumor (Nowak and Sigmund 1998), ability to choose the opponent (Ashlock et al. 1996; Cooper and Wallace 1998), variable investment (Roberts and Sherratt 1998), negotiation (McNamara et al. 1999), and so on (for review, see Brembs 1996). Most of those factors assume additional environmental conditions to the pure Prisoners' Dilemma game. When players must play a pure Prisoners' Dilemma game, the only available information is history of games.

We showed that robust cooperation can be maintained without additional assumption on environment if players have a very primitive ability to learn from history of games.

There are some studies in which players in Prisoner's Dilemma are modeled as learning automata (Rubinstein 1986; Billard 1995, 1996; Harrald and Fogel 1995). These models can be viewed as complicated variants of reinforcement learners. Automata or a neural network is a model of brain, and the study of behavior of players with such complicated learning may be important for analysis of higher animals. However, we go to the opposite. We are interested in the behavior of the most simple form of reinforcement learner, which seems to be provided with many animals. Our result shows that players with such a simple and intuitively familiar learning rule establish very efficient and robust cooperation.

Acknowledgments We thank R. Harrison for valuable comments and improving the English. We also thank M. Higashi for helpful suggestions. We thank an anonymous referee for valuable comments. We used gcc on Linux for computer simulations and we thank all developers and programmers who support Linux, the GNU project, and other useful freeware. This work is partly supported by a Grant-in-Aid for Scientific Research from the Japan Ministry of Education, Science, Sports and Culture (Creative Basic Research Program: DIVER).

References

Ashlock D, Smucker MD, Stanley EA, Tesfatsion L (1996) Preferential partner selection in an evolutionary study of Prisoner's Dilemma. Biosystems 37:99–125

Axelrod R (1984) The evolution of cooperation. Basic Books, New York

Axelrod R, Hamilton WD (1981) The evolution of cooperation. Science 211:1390–1396

Billard EA (1995) Adaptation in a stochastic prisoner's dilemma with delayed information. Biosystems 37:211–227

Billard EA (1996) Evolutionary strategies of stochastic learning automata in the prisoner's dilemma. Biosystems 39:93–107

Boyd R, Lorberbaum JP (1987) No pure strategy is evolutionarily stable in the repeated Prisoner's Dilemma game. Nature (Lond) 327:58–59

Brattstrom BH (1990) Maze learning in the fire-bellied toad, *Bombina orientalis*. J Herpetol 24:44–47

Brauchli K, Killingback T, Doebeli M (1999) Evolution of cooperation in spatially structured populations. J Theor Biol 200:405–417

Brembs B (1996) Chaos, cheating and cooperation: potential solutions to the Prisoner's Dilemma. Oikos 76:14–24

Cooper B, Wallace C (1998) Evolution, partnerships and cooperation. J Theor Biol 195:315–328

Ferriere R, Michod RE (1996) The evolution of cooperation in spatially heterogeneous populations. Am Nat 147:692–717

Harrald PG, Fogel DB (1995) Evolving continuous behaviors in the Iterated Prisoner's Dilemma. Biosystems 37:135–145

Lindgren K, Nordahl MG (1994) Evolutionary dynamics of spatial games. Physica D 75:292–309

Lorberbaum JP (1994) No strategy is evolutionarily stable in the repeated Prisoner's Dilemma. J Theor Biol 168:117–130

Maynard-Smith J, Price GR (1973) The logic of animal conflict. Nature (Lond) 246:15–18

McNamara JM, Gasson CE, Houston A (1999) Incorporating rules for responding into evolutionary games. Nature (Lond) 401:368–371

Nowak MA, Sigmund K (1993a) A strategy of win-stay lose-shift that outperforms tit-for-tat in the Prisoner's Dilemma game. Nature (Lond) 364:56–58

Nowak MA, Sigmund K (1993b) Chaos and the evolution of cooperation. Proc Natl Acad Sci USA 90:5091–5094

Nowak MA, Sigmund K (1998) Evolution of indirect reciprocity by image scoring. Nature (Lond) 393:573–577

Papini MR (1997) Role of reinforcement in spaced-trial operant learning in pigeons (*Columba livia*). J Comp Psychol 111:275–285

Posch M (1999) Win-stay lose-shift strategies for repeated games – memory length, aspiration levels and noise. J Theor Biol 198:183–195

Roberts G, Sherratt TN (1998) Development of cooperative relationships through increasing investment. Nature (Lond) 394:175–179

Rubinstein A (1986) Finite automata play the repeated prisoner's dilemma. J Econ Theory 39:83–96

Sandholm TW, Crites RH (1995) Multiagent reinforcement in the Iterated Prisoner's Dilemma. Biosystems 37:147–166

Stephens DW, Clements KC (1998) Game theory and learning. In: Dugatkin LA, Reeve HK (eds) Game theory and animal behavior. Oxford University Press, New York, pp 239–250

Stephens DW, Nishimura K, Toyer KB (1995) Error and discounting in the Iterated Prisoner's Dilemma. J Theor Biol 176:457–469

Whitehouse MEA (1997) Experience influences male-male contests in the spider *Argyrodes antipodiana* (Theridiidae: Araneae). Anim Behav 53:913–923