Symmetric equilibrium of multi-agent reinforcement learning in repeated prisoner's dilemma

Yuki Usui

Faculty of Science, Yamaguchi University, Yamaguchi 753-8511, Japan

Masahiko Ueda

Graduate School of Sciences and Technology for Innovation, Yamaguchi University, Yamaguchi 753-8511, Japan

Abstract

We investigate the repeated prisoner's dilemma game where both players alternately use reinforcement learning to obtain their optimal memory-one strategies. We theoretically solve the simultaneous Bellman optimality equations of reinforcement learning. We find that the Win-stay Lose-shift strategy, the Grim strategy, and the strategy which always defects can form symmetric equilibrium of the mutual reinforcement learning process amongst all deterministic memory-one strategies.

Keywords: Repeated prisoner's dilemma game; Reinforcement learning

1. Introduction

The prisoner's dilemma game describes a dilemma where rational behavior of each player cannot achieve a favorable situation for both players [1]. In the game, each player chooses cooperation or defection. Each player can obtain more payoff by taking defection than by taking cooperation regardless of the opponent's action. Then, mutual defection is realized as a result of rational thought of both players, while payoffs of both players increase when both players choose cooperation. Although the Nash equilibrium of the one-shot game is mutual

Email addresses: i007de@yamaguchi-u.ac.jp (Yuki Usui), m.ueda@yamaguchi-u.ac.jp (Masahiko Ueda)

defection, when the game is infinitely repeated, it has been known that mutual cooperation can be achieved as the Nash equilibrium. This fact is known as the folk theorem. Because the repeated version of the prisoner's dilemma game is also simple, it has substantially been investigated [2].

Recently, reinforcement learning technique attracts much attentions in the context of game theory [3, 4, 5, 6, 7, 8, 9, 10, 11]. In reinforcement learning, a player gradually learns his/her optimal strategy against his/her opponents. Both learning by a single player and learning by several players have been investigated. Because rationality of players is bounded in reality, modeling of players as learning agents is crucial [12]. It is also significant in the context of reinforcement learning, since the original reinforcement learning was formulated for Markov decision process with stationary environments [13]. Because the existence of multiple agents in game theory leads to non-stationarity of environments for each player, the standard application of reinforcement learning to games breaks down [4, 14], and further theoretical understanding of reinforcement learning in game theory is needed. Moreover, since the acquisition process of optimal strategies in reinforcement learning is generally different from that in evolutionary game theory [15], accumulating knowledge about equilibrium in each learning dynamics is needed.

In this paper, we investigate the situation where both players alternately learn their optimal strategies by using reinforcement learning in the repeated prisoner's dilemma game. We theoretically derive equilibrium points of mutual reinforcement learning where both players take the same deterministic strategy. We find that the strategy which always defects (All-D), the Win-stay Lose-Shift (WSLS) strategy [16], and the Grim strategy can form such symmetric equilibrium amongst all memory-one deterministic strategies.

This paper is organized as follows. In Section 2, we introduce the repeated prisoner's dilemma game, and players using reinforcement learning. In Section 3, we theoretically derive deterministic optimal strategies against the strategy of a learning opponent. In Section 4, we provide numerical results by using Q-learning which support our theoretical results. Section 5 is devoted to con-

clusion.

2. Model

We consider the repeated prisoner's dilemma game [3]. There are two players in the game, and each player is described as 1 and 2. Each player chooses cooperation (C) or defection (D) on every trial. The action of player a is written as $\sigma_a \in \{C, D\}$, and we collectively write $\boldsymbol{\sigma} := (\sigma_1, \sigma_2)$. The payoff of player $a \in \{1, 2\}$ when the state is $\boldsymbol{\sigma}$ is described as $r_a(\boldsymbol{\sigma})$. The payoffs in the prisoner's dilemma game are defined as

$$\begin{pmatrix} r_1(C,C), r_2(C,C) & r_1(C,D), r_2(C,D) \\ r_1(D,C), r_2(D,C) & r_1(D,D), r_2(D,D) \end{pmatrix} = \begin{pmatrix} R,R & S,T \\ T,S & P,P \end{pmatrix} \quad (1)$$

with $T > R > P > S$ and $2R > T + S$. We consider the situation where both players use memory-one strategies. The memory-one strategy of player a is described as the conditional probability $T_a(\sigma_a|\boldsymbol{\sigma}')$ of taking action σ_a when the state in the previous round is $\boldsymbol{\sigma}'$. (In this paper, we investigate only the game with perfect monitoring, where players can perfectly observe the actions of both players in the previous round.) Then, when we define the probability distribution of a state $\boldsymbol{\sigma}'$ at time t by $P(\boldsymbol{\sigma}', t)$, the time evolution of this system is described as the Markov chain

$$P(\boldsymbol{\sigma}, t+1) = \sum_{\boldsymbol{\sigma}'} T(\boldsymbol{\sigma}|\boldsymbol{\sigma}') P(\boldsymbol{\sigma}', t) \quad (2)$$

with the transition probability

$$T(\boldsymbol{\sigma}|\boldsymbol{\sigma}') := \prod_{a=1}^{2} T_a(\sigma_a|\boldsymbol{\sigma}'). \quad (3)$$

Below we introduce the notation $-a := \{1, 2\} \backslash a$.

We consider the situation that both players learn their strategies by reinforcement learning [13]. We assume that two players alternately learn and update their strategies [17], that is, player 1 first learns her strategy against a fixed initial strategy of player 2, then player 2 learns his strategy against

the strategy of player 1, then player 1 learns her strategy against the strategy of player 2, and so on. In other words, the two players infinitely repeat the infinitely repeated game, and their strategies are updated after each repeated game is played and their long-term payoffs are calculated. We assume that the strategy of player 1 is updated in n-th game with $n = 2m - 1$ ($m \in \mathbb{N}$) and the strategy of player 2 is updated in n-th game with $n = 2m$ ($m \in \mathbb{N}$). We write the strategies of player a at n-th game as $T_a^{(n)}(\sigma_a|\boldsymbol{\sigma}')$.

In reinforcement learning, each player learns mapping (called policy) from a state to his/her action so as to maximize his/her expected future reward. In our memory-one situation, a state and an action of player a are regarded as the state $\boldsymbol{\sigma}'$ in the previous round and the action σ_a in the present round, respectively. We define the action-value function of player a as

$$Q_a\left(\sigma_a^{(1)}, \boldsymbol{\sigma}^{(0)}\right) := \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k r_a(t+k+1) \middle| \sigma_a(t+1) = \sigma_a^{(1)}, \boldsymbol{\sigma}(t) = \boldsymbol{\sigma}^{(0)}\right], \tag{4}$$

where γ is a discounting factor satisfying $0 \leq \gamma < 1$. The action $\sigma_a(t)$ represents the action of player a at round t. Similarly, the payoff $r_a(t)$ represents the payoff of player a at round t, that is, $r_a(t) := r_a(\boldsymbol{\sigma}(t))$. Due to the Markov property, the action-value function Q obeys the Bellman equation against a fixed strategy T_{-a} of the opponent:

$$\begin{aligned} Q_a\left(\sigma_a^{(1)}, \boldsymbol{\sigma}^{(0)}\right) &= \sum_{\sigma_{-a}^{(1)}} T_{-a}\left(\sigma_{-a}^{(1)}|\boldsymbol{\sigma}^{(0)}\right) r_a\left(\boldsymbol{\sigma}^{(1)}\right) \\ &\quad + \gamma \sum_{\sigma_a^{(2)}} \sum_{\sigma_{-a}^{(1)}} T_a\left(\sigma_a^{(2)}|\boldsymbol{\sigma}^{(1)}\right) T_{-a}\left(\sigma_{-a}^{(1)}|\boldsymbol{\sigma}^{(0)}\right) Q_a\left(\sigma_a^{(2)}, \boldsymbol{\sigma}^{(1)}\right). \end{aligned} \tag{5}$$

It has been known that the optimal value of Q obeys the following Bellman optimality equation:

$$Q_a^*\left(\sigma_a^{(1)}, \boldsymbol{\sigma}^{(0)}\right) = \sum_{\sigma_{-a}^{(1)}} T_{-a}\left(\sigma_{-a}^{(1)}|\boldsymbol{\sigma}^{(0)}\right) r_a\left(\boldsymbol{\sigma}^{(1)}\right) + \gamma \sum_{\sigma_{-a}^{(1)}} T_{-a}\left(\sigma_{-a}^{(1)}|\boldsymbol{\sigma}^{(0)}\right) \max_{\sigma_a^{(2)}} Q_a^*\left(\sigma_a^{(2)}, \boldsymbol{\sigma}^{(1)}\right) \tag{6}$$

with the support

$$\mathrm{supp}T_a\left(\cdot|\,\boldsymbol{\sigma}^{(0)}\right) \quad = \quad \arg\max_{\sigma} Q_a^*\left(\sigma, \boldsymbol{\sigma}^{(0)}\right). \tag{7}$$

In other words, in the optimal policy against T_{-a}, player a takes the action σ_a which maximizes the value of $Q_a^*\left(\cdot, \boldsymbol{\sigma}^{(0)}\right)$ when the state at the previous round is $\boldsymbol{\sigma}^{(0)}$.

In sum, in the $(2m-1)$-th game, player 1 learns $T_1^{(2m-1)}\left(\sigma_1|\boldsymbol{\sigma'}\right)$ against $T_2^{(2m-2)}\left(\sigma_2|\boldsymbol{\sigma'}\right)$ by calculating $Q_1^{*(2m-1)}\left(\sigma, \boldsymbol{\sigma}^{(0)}\right)$, where $Q_a^{*(n)}\left(\sigma, \boldsymbol{\sigma}^{(0)}\right)$ represents the optimal action-value function of player a in the n-th game. In the $2m$-th game, player 2 learns $T_2^{(2m)}\left(\sigma_2|\boldsymbol{\sigma'}\right)$ against $T_1^{(2m-1)}\left(\sigma_1|\boldsymbol{\sigma'}\right)$ by calculating $Q_2^{*(2m)}\left(\sigma, \boldsymbol{\sigma}^{(0)}\right)$. We are interested in the fixed points of the dynamics, that is, $T_a^{(\infty)}\left(\sigma_a|\boldsymbol{\sigma'}\right)$ and $Q_a^{*(\infty)}\left(\sigma, \boldsymbol{\sigma}^{(0)}\right)$.

In this paper, we investigate only situations that the support (7) contains only one action, that is, we investigate only deterministic strategies. Because the number of deterministic memory-one strategies in the repeated prisoner's dilemma game is sixteen, we check whether each deterministic strategy forms equilibrium or not.

3. Results

We consider symmetric solutions of Eq. (6), that is,

$$Q_1^*\left(\sigma_1^{(1)}, \boldsymbol{\sigma}^{(0)}\right) \quad = \quad \sum_{\sigma_2^{(1)}} T_2\left(\sigma_2^{(1)}|\boldsymbol{\sigma}^{(0)}\right) r_1\left(\boldsymbol{\sigma}^{(1)}\right) + \gamma \sum_{\sigma_2^{(1)}} T_2\left(\sigma_2^{(1)}|\boldsymbol{\sigma}^{(0)}\right) \max_{\sigma_1^{(2)}} Q_1^*\left(\sigma_1^{(2)}, \boldsymbol{\sigma}^{(1)}\right) \tag{8}$$

with

$$T_2\left(C|C,C\right) \quad = \quad \mathbb{I}\left(Q_1^*(C,C,C) > Q_1^*(D,C,C)\right) \tag{9}$$

$$T_2\left(C|C,D\right) \quad = \quad \mathbb{I}\left(Q_1^*(C,D,C) > Q_1^*(D,D,C)\right) \tag{10}$$

$$T_2\left(C|D,C\right) \quad = \quad \mathbb{I}\left(Q_1^*(C,C,D) > Q_1^*(D,C,D)\right) \tag{11}$$

$$T_2\left(C|D,D\right) \quad = \quad \mathbb{I}\left(Q_1^*(C,D,D) > Q_1^*(D,D,D)\right) \tag{12}$$

where $\mathbb{I}(\cdots)$ is the indicator function that returns 1 when $\cdots$ holds and 0 otherwise. Then, Eq. (6) becomes

$$\begin{aligned} Q_1^*(C,C,C) &= \mathbb{I}\left(Q_1^*(C,C,C) > Q_1^*(D,C,C)\right)\left\{R + \gamma \max_\sigma Q_1^*(\sigma,C,C)\right\} \\ &\quad + \mathbb{I}\left(Q_1^*(C,C,C) < Q_1^*(D,C,C)\right)\left\{S + \gamma \max_\sigma Q_1^*(\sigma,C,D)\right\} \end{aligned} \tag{13}$$

$$\begin{aligned} Q_1^*(C,C,D) &= \mathbb{I}\left(Q_1^*(C,D,C) > Q_1^*(D,D,C)\right)\left\{R + \gamma \max_\sigma Q_1^*(\sigma,C,C)\right\} \\ &\quad + \mathbb{I}\left(Q_1^*(C,D,C) < Q_1^*(D,D,C)\right)\left\{S + \gamma \max_\sigma Q_1^*(\sigma,C,D)\right\} \end{aligned} \tag{14}$$

$$\begin{aligned} Q_1^*(C,D,C) &= \mathbb{I}\left(Q_1^*(C,C,D) > Q_1^*(D,C,D)\right)\left\{R + \gamma \max_\sigma Q_1^*(\sigma,C,C)\right\} \\ &\quad + \mathbb{I}\left(Q_1^*(C,C,D) < Q_1^*(D,C,D)\right)\left\{S + \gamma \max_\sigma Q_1^*(\sigma,C,D)\right\} \end{aligned} \tag{15}$$

$$\begin{aligned} Q_1^*(C,D,D) &= \mathbb{I}\left(Q_1^*(C,D,D) > Q_1^*(D,D,D)\right)\left\{R + \gamma \max_\sigma Q_1^*(\sigma,C,C)\right\} \\ &\quad + \mathbb{I}\left(Q_1^*(C,D,D) < Q_1^*(D,D,D)\right)\left\{S + \gamma \max_\sigma Q_1^*(\sigma,C,D)\right\} \end{aligned} \tag{16}$$

$$\begin{aligned} Q_1^*(D,C,C) &= \mathbb{I}\left(Q_1^*(C,C,C) > Q_1^*(D,C,C)\right)\left\{T + \gamma \max_\sigma Q_1^*(\sigma,D,C)\right\} \\ &\quad + \mathbb{I}\left(Q_1^*(C,C,C) < Q_1^*(D,C,C)\right)\left\{P + \gamma \max_\sigma Q_1^*(\sigma,D,D)\right\} \end{aligned} \tag{17}$$

$$\begin{aligned} Q_1^*(D,C,D) &= \mathbb{I}\left(Q_1^*(C,D,C) > Q_1^*(D,D,C)\right)\left\{T + \gamma \max_\sigma Q_1^*(\sigma,D,C)\right\} \\ &\quad + \mathbb{I}\left(Q_1^*(C,D,C) < Q_1^*(D,D,C)\right)\left\{P + \gamma \max_\sigma Q_1^*(\sigma,D,D)\right\} \end{aligned} \tag{18}$$

$$\begin{aligned} Q_1^*(D,D,C) &= \mathbb{I}\left(Q_1^*(C,C,D) > Q_1^*(D,C,D)\right)\left\{T + \gamma \max_\sigma Q_1^*(\sigma,D,C)\right\} \\ &\quad + \mathbb{I}\left(Q_1^*(C,C,D) < Q_1^*(D,C,D)\right)\left\{P + \gamma \max_\sigma Q_1^*(\sigma,D,D)\right\} \end{aligned} \tag{19}$$

$$
\begin{aligned}
Q_1^*(D,D,D) &= \mathbb{I}(Q_1^*(C,D,D) > Q_1^*(D,D,D)) \left\{ T + \gamma \max_\sigma Q_1^*(\sigma,D,C) \right\} \\
&\quad + \mathbb{I}(Q_1^*(C,D,D) < Q_1^*(D,D,D)) \left\{ P + \gamma \max_\sigma Q_1^*(\sigma,D,D) \right\}.
\end{aligned} \tag{20}
$$

For simplicity, we introduce the following notation:

$$
\begin{aligned}
q_1 &:= Q_1^*(C,C,C) \\
q_2 &:= Q_1^*(C,C,D) \\
q_3 &:= Q_1^*(C,D,C) \\
q_4 &:= Q_1^*(C,D,D) \\
q_5 &:= Q_1^*(D,C,C) \\
q_6 &:= Q_1^*(D,C,D) \\
q_7 &:= Q_1^*(D,D,C) \\
q_8 &:= Q_1^*(D,D,D).
\end{aligned} \tag{21}
$$

We consider the following sixteen situations separately.

3.1. Case 1: $q_1 > q_5$, $q_2 > q_6$, $q_3 > q_7$, *and* $q_4 > q_8$

For this case, the strategy obtained by reinforcement learning is the All-C strategy. The solution of Eq. (6) is

$$
q_1 = q_2 = q_3 = q_4 = \frac{1}{1-\gamma} R \tag{22}
$$

$$
q_5 = q_6 = q_7 = q_8 = T + \frac{\gamma}{1-\gamma} R. \tag{23}
$$

This contradicts with the definition of the game $T > R$.

3.2. Case 2: $q_1 > q_5$, $q_2 > q_6$, $q_3 > q_7$, *and* $q_4 < q_8$

The solution of Eq. (6) is

$$
q_1 = q_2 = q_3 = \frac{1}{1-\gamma} R \tag{24}
$$

$$
q_4 = S + \frac{\gamma}{1-\gamma} R \tag{25}
$$

$$
q_5 = q_6 = q_7 = T + \frac{\gamma}{1-\gamma} R \tag{26}
$$

$$
q_8 = \frac{1}{1-\gamma} P. \tag{27}
$$

This contradicts with the definition of the game $T > R$.

3.3. Case 3: $q_1 > q_5$, $q_2 > q_6$, $q_3 < q_7$, *and* $q_4 > q_8$

The solution of Eq. (6) is

$$q_1 = q_3 = q_4 = \frac{1}{1-\gamma}R \tag{28}$$

$$q_2 = \frac{1}{1-\gamma}S \tag{29}$$

$$q_5 = q_7 = q_8 = \frac{1}{1-\gamma}T \tag{30}$$

$$q_6 = P + \frac{\gamma}{1-\gamma}R. \tag{31}$$

This contradicts with the definition of the game $T > R$.

3.4. Case 4: $q_1 > q_5$, $q_2 > q_6$, $q_3 < q_7$, *and* $q_4 < q_8$

For this case, the strategy obtained by reinforcement learning is "Repeat" [18]. The solution of Eq. (6) is

$$q_1 = q_3 = \frac{1}{1-\gamma}R \tag{32}$$

$$q_2 = q_4 = \frac{1}{1-\gamma}S \tag{33}$$

$$q_5 = q_7 = \frac{1}{1-\gamma}T \tag{34}$$

$$q_6 = q_8 = \frac{1}{1-\gamma}P. \tag{35}$$

This contradicts with the definition of the game $T > R$.

3.5. Case 5: $q_1 > q_5$, $q_2 < q_6$, $q_3 > q_7$, *and* $q_4 > q_8$

The solution of Eq. (6) is

$$q_1 = q_2 = q_4 = \frac{1}{1-\gamma}R \tag{36}$$

$$q_3 = \frac{1}{1-\gamma^2}S + \frac{\gamma}{1-\gamma^2}T \tag{37}$$

$$q_5 = q_6 = q_8 = \frac{1}{1-\gamma^2}T + \frac{\gamma}{1-\gamma^2}S \tag{38}$$

$$q_7 = P + \frac{\gamma}{1-\gamma}R. \tag{39}$$

This contradicts with $2R > T + S$.

3.6. Case 6: $q_1 > q_5$, $q_2 < q_6$, $q_3 > q_7$, *and* $q_4 < q_8$

For this case, the strategy obtained by reinforcement learning is Tit-for-Tat (TFT) [1, 19]. The solution of Eq. (6) is

$$q_1 = q_2 = \frac{1}{1-\gamma}R \tag{40}$$

$$q_3 = q_4 = \frac{1}{1-\gamma^2}S + \frac{\gamma}{1-\gamma^2}T \tag{41}$$

$$q_5 = q_6 = \frac{1}{1-\gamma^2}T + \frac{\gamma}{1-\gamma^2}S \tag{42}$$

$$q_7 = q_8 = \frac{1}{1-\gamma}P. \tag{43}$$

This solution becomes consistent with the condition of the case only when $T + S = R + P$ and $\gamma = \frac{T-R}{R-S}$.

3.7. Case 7: $q_1 > q_5$, $q_2 < q_6$, $q_3 < q_7$, *and* $q_4 > q_8$

For this case, the strategy obtained by reinforcement learning is Win-stay-Lose-Shift (WSLS) [16]. The solution of Eq. (6) is

$$q_1 = q_4 = \frac{1}{1-\gamma}R \tag{44}$$

$$q_2 = q_3 = S + \gamma P + \frac{\gamma^2}{1-\gamma}R \tag{45}$$

$$q_5 = q_8 = T + \gamma P + \frac{\gamma^2}{1-\gamma}R \tag{46}$$

$$q_6 = q_7 = P + \frac{\gamma}{1-\gamma}R. \tag{47}$$

This solution becomes consistent with the condition of the case when $T+P < 2R$ and $\gamma > \frac{T-R}{R-P}$.

3.8. Case 8: $q_1 > q_5$, $q_2 < q_6$, $q_3 < q_7$, *and* $q_4 < q_8$

For this case, the strategy obtained by reinforcement learning is the Grim strategy. The solution of Eq. (6) is

$$q_1 = \frac{1}{1-\gamma}R \tag{48}$$

$$q_2 = q_3 = q_4 = S + \frac{\gamma}{1-\gamma}P \tag{49}$$

$$q_5 = T + \frac{\gamma}{1-\gamma}P \tag{50}$$

$$q_6 = q_7 = q_8 = \frac{1}{1-\gamma}P. \tag{51}$$

This solution becomes consistent with the condition of the case when $\gamma > \frac{T-R}{T-P}$.

3.9. Case 9: $q_1 < q_5$, $q_2 > q_6$, $q_3 > q_7$, *and* $q_4 > q_8$

For this case, the strategy obtained by reinforcement learning is the anti-Grim strategy. The solution of Eq. (6) is

$$q_1 = S + \frac{\gamma}{1-\gamma^2}R + \frac{\gamma^2}{1-\gamma^2}P \tag{52}$$

$$q_2 = q_3 = q_4 = \frac{1}{1-\gamma^2}R + \frac{\gamma}{1-\gamma^2}P \tag{53}$$

$$q_5 = \frac{1}{1-\gamma^2}P + \frac{\gamma}{1-\gamma^2}R \tag{54}$$

$$q_6 = q_7 = q_8 = T + \frac{\gamma}{1-\gamma^2}R + \frac{\gamma^2}{1-\gamma^2}P. \tag{55}$$

This contradicts with $\gamma \geq 0$.

3.10. Case 10: $q_1 < q_5$, $q_2 > q_6$, $q_3 > q_7$, *and* $q_4 < q_8$

For this case, the strategy obtained by reinforcement learning is anti-Win-stay-Lose-Shift (AWSLS). The solution of Eq. (6) is

$$q_1 = q_4 = S + \gamma R + \frac{\gamma^2}{1-\gamma}P \tag{56}$$

$$q_2 = q_3 = R + \frac{\gamma}{1-\gamma}P \tag{57}$$

$$q_5 = q_8 = \frac{1}{1-\gamma}P \tag{58}$$

$$q_6 = q_7 = T + \gamma R + \frac{\gamma^2}{1-\gamma}P. \tag{59}$$

This contradicts with $\gamma \geq 0$.

3.11. Case 11: $q_1 < q_5$, $q_2 > q_6$, $q_3 < q_7$, *and* $q_4 > q_8$

For this case, the strategy obtained by reinforcement learning is anti-Tit-for-Tat (ATFT). The solution of Eq. (6) is

$$q_1 = q_2 = \frac{1}{1-\gamma} S \tag{60}$$

$$q_3 = q_4 = \frac{1}{1-\gamma^2} R + \frac{\gamma}{1-\gamma^2} P \tag{61}$$

$$q_5 = q_6 = \frac{1}{1-\gamma^2} P + \frac{\gamma}{1-\gamma^2} R \tag{62}$$

$$q_7 = q_8 = \frac{1}{1-\gamma} T. \tag{63}$$

This contradicts with $\gamma \geq 0$.

3.12. Case 12: $q_1 < q_5$, $q_2 > q_6$, $q_3 < q_7$, *and* $q_4 < q_8$

The solution of Eq. (6) is

$$q_1 = q_2 = q_4 = \frac{1}{1-\gamma} S \tag{64}$$

$$q_3 = R + \frac{\gamma}{1-\gamma} P \tag{65}$$

$$q_5 = q_6 = q_8 = \frac{1}{1-\gamma} P \tag{66}$$

$$q_7 = \frac{1}{1-\gamma} T. \tag{67}$$

This contradicts with the definition of the game $P > S$.

3.13. Case 13: $q_1 < q_5$, $q_2 < q_6$, $q_3 > q_7$, *and* $q_4 > q_8$

For this case, the strategy obtained by reinforcement learning is anti-Repeat. The solution of Eq. (6) is

$$q_1 = q_3 = \frac{1}{1-\gamma^2} S + \frac{\gamma}{1-\gamma^2} T \tag{68}$$

$$q_2 = q_4 = \frac{1}{1-\gamma^2} R + \frac{\gamma}{1-\gamma^2} P \tag{69}$$

$$q_5 = q_7 = \frac{1}{1-\gamma^2} P + \frac{\gamma}{1-\gamma^2} R \tag{70}$$

$$q_6 = q_8 = \frac{1}{1-\gamma^2} T + \frac{\gamma}{1-\gamma^2} S. \tag{71}$$

This solution becomes consistent with the condition of the case only when $T + S = R + P$ and $\gamma = 1$.

3.14. Case 14: $q_1 < q_5$, $q_2 < q_6$, $q_3 > q_7$, *and* $q_4 < q_8$

The solution of Eq. (6) is

$$q_1 = q_3 = q_4 = \frac{1}{1-\gamma^2} S + \frac{\gamma}{1-\gamma^2} T \tag{72}$$

$$q_2 = R + \frac{\gamma}{1-\gamma} P \tag{73}$$

$$q_5 = q_7 = q_8 = \frac{1}{1-\gamma} P \tag{74}$$

$$q_6 = \frac{1}{1-\gamma^2} T + \frac{\gamma}{1-\gamma^2} S. \tag{75}$$

This solution becomes consistent with the condition of the case only when $T + S > 2P$ and $\gamma = \frac{P-S}{T-S}$.

3.15. Case 15: $q_1 < q_5$, $q_2 < q_6$, $q_3 < q_7$, *and* $q_4 > q_8$

The solution of Eq. (6) is

$$q_1 = q_2 = q_3 = S + \frac{\gamma}{1-\gamma^2} P + \frac{\gamma^2}{1-\gamma^2} R \tag{76}$$

$$q_4 = \frac{1}{1-\gamma^2} R + \frac{\gamma}{1-\gamma^2} P \tag{77}$$

$$q_5 = q_6 = q_7 = \frac{1}{1-\gamma^2} P + \frac{\gamma}{1-\gamma^2} R \tag{78}$$

$$q_8 = T + \frac{\gamma}{1-\gamma^2} P + \frac{\gamma^2}{1-\gamma^2} R. \tag{79}$$

This contradicts with the definition of the game $T > R$.

3.16. Case 16: $q_1 < q_5$, $q_2 < q_6$, $q_3 < q_7$, *and* $q_4 < q_8$

For this case, the strategy obtained by reinforcement learning is the All-D strategy. The solution of Eq. (6) is

$$q_1 = q_2 = q_3 = q_4 = S + \frac{\gamma}{1-\gamma} P \tag{80}$$

$$q_5 = q_6 = q_7 = q_8 = \frac{1}{1-\gamma} P. \tag{81}$$

This solution is always consistent with the condition of the case.

number	$q_1 \lessgtr q_5$	$q_2 \lessgtr q_6$	$q_3 \lessgtr q_7$	$q_4 \lessgtr q_8$	strategy $\boldsymbol{T}_1(C)$	name	Equilibrium?
Case 1	$>$	$>$	$>$	$>$	$(1,1,1,1)^{\mathsf{T}}$	All-C	No
Case 2	$>$	$>$	$>$	$<$	$(1,1,1,0)^{\mathsf{T}}$		No
Case 3	$>$	$>$	$<$	$>$	$(1,1,0,1)^{\mathsf{T}}$		No
Case 4	$>$	$>$	$<$	$<$	$(1,1,0,0)^{\mathsf{T}}$	Repeat	No
Case 5	$>$	$<$	$>$	$>$	$(1,0,1,1)^{\mathsf{T}}$		No
Case 6	$>$	$<$	$>$	$<$	$(1,0,1,0)^{\mathsf{T}}$	TFT	No in general
Case 7	$>$	$<$	$<$	$>$	$(1,0,0,1)^{\mathsf{T}}$	WSLS	Yes for $\gamma > \frac{T-R}{R-P}$
Case 8	$>$	$<$	$<$	$<$	$(1,0,0,0)^{\mathsf{T}}$	Grim	Yes for $\gamma > \frac{T-R}{T-P}$
Case 9	$<$	$>$	$>$	$>$	$(0,1,1,1)^{\mathsf{T}}$	anti-Grim	No
Case 10	$<$	$>$	$>$	$<$	$(0,1,1,0)^{\mathsf{T}}$	AWSLS	No
Case 11	$<$	$>$	$<$	$>$	$(0,1,0,1)^{\mathsf{T}}$	ATFT	No
Case 12	$<$	$>$	$<$	$<$	$(0,1,0,0)^{\mathsf{T}}$		No
Case 13	$<$	$<$	$>$	$>$	$(0,0,1,1)^{\mathsf{T}}$	anti-Repeat	No in general
Case 14	$<$	$<$	$>$	$<$	$(0,0,1,0)^{\mathsf{T}}$		No in general
Case 15	$<$	$<$	$<$	$>$	$(0,0,0,1)^{\mathsf{T}}$		No
Case 16	$<$	$<$	$<$	$<$	$(0,0,0,0)^{\mathsf{T}}$	All-D	Yes

Table 1: Summary of the results.

3.17. Summary

From the above subsections, we find that the symmetric solution of the Bellman optimality equation exists in finite regions of the parameter γ only for the case 7, 8, and 16. In other words, only WSLS, the Grim strategy, and the All-D strategy can form the symmetric equilibrium of mutual reinforcement learning. TFT does not form symmetric equilibrium. (The optimal strategy against TFT is investigated in detail in Appendix A.) The results are summarized in Table 1, where the strategy vector of player 1 is defined by

$$\boldsymbol{T}_1(C) \;:=\; \begin{pmatrix} T_1\left(C|C,C\right) \\ T_1\left(C|C,D\right) \\ T_1\left(C|D,C\right) \\ T_1\left(C|D,D\right) \end{pmatrix}. \tag{82}$$

The reason why TFT does not form equilibrium can be intuitively explained as follows. We consider the situation that player 2 uses TFT, and the previous

state was (D, C). By definition, the action-value function (4) represents expected future reward when the previous state was $\boldsymbol{\sigma}^{(0)}$. If the strategy of player 1 is also TFT, the sequence

$$(D, C) \to (C, D) \to (D, C) \to (C, D) \to \cdots \tag{83}$$

is realized. If the strategy of player 1 is All-C, the sequence

$$(D, C) \to (C, D) \to (C, C) \to (C, C) \to \cdots \tag{84}$$

is realized. Because of $T + S < 2R$, the latter results in larger total payoff than the former. Therefore, as explained in Appendix A, the optimal strategy against TFT is not TFT.

One may recall that All-D, WSLS and Grim form subgame perfect equilibria in the repeated prisoner's dilemma game, but TFT dose not [20]. Therefore, our reinforcement learning equilibrium seems to be similar to subgame perfect equilibrium. In fact, the above discussion that TFT does not form reinforcement learning equilibrium is similar to the discussion that TFT does not form subgame perfect equilibrium. However, we expect that reinforcement learning equilibrium is weaker than subgame perfect equilibrium, since, in the definition of subgame perfect equilibrium, arbitrary histories are considered. Relation between them should be clarified in future.

4. Numerical results

In this section, we check the theoretical results in the previous section by numerical simulation. We use Q-learning [13] as a method of reinforcement learning. In Q-learning, the optimal action-value function of the agent a against a fixed strategy of the agent $-a$ is learned through the following update rule:

$$Q_a^{(t+1)}\left(\sigma_a^{(1)}, \boldsymbol{\sigma}^{(0)}\right) = Q_a^{(t)}\left(\sigma_a^{(1)}, \boldsymbol{\sigma}^{(0)}\right) + \eta\left(r_a + \gamma \max_{\sigma_a^{(2)}} Q_a^{(t)}\left(\sigma_a^{(2)}, \boldsymbol{\sigma}^{(1)}\right) - Q_a^{(t)}\left(\sigma_a^{(1)}, \boldsymbol{\sigma}^{(0)}\right)\right), \tag{85}$$

where r_a is the reward by taking action $\sigma_a^{(1)}$ when the state is $\boldsymbol{\sigma}^{(0)}$, and $\boldsymbol{\sigma}^{(1)}$ is the next state. The parameter η is called the learning rate. Here, we assume

that, in each step, the agent a chooses the action $\sigma_a^{(1)}$ by using ϵ-greedy search, that is, the agent a chooses an action uniformly randomly among all possible actions with probability ϵ, and chooses the best action with respect to the current action-value function with probability $1-\epsilon$. As before, we consider the situation that two agents alternately learn their optimal strategies until Q values converge.

We set parameters $(R,S,T,P)=(4,0,6,1)$, $\eta=0.2$, and $\epsilon=0.01$. In the numarical calculation of Q, we take the statistical average over 10^3 realizations. The initial condition of Q is $Q\left(\sigma^{(1)},\boldsymbol{\sigma}^{(0)}\right)=0$ for all $\sigma^{(1)}$ and $\boldsymbol{\sigma}^{(0)}$.

In Figure 1, we display the time evolution of Q_1 when the strategy of player 2 is WSLS. According to Appendix A, the optimal strategy against WSLS is WSLS for $\gamma > (T-R)/(R-P)$ and All-D for $\gamma < (T-R)/(R-P)$. On the top panel of Figure 1, we provide the numerical results for $\gamma = 0.9 > (T-R)/(R-P)=2/3$. The theoretical value of Q_1 is also provided in Appendix A:

$$q_1 = q_4 = \frac{1}{1-\gamma}R = 40 \tag{86}$$

$$q_2 = q_3 = S+\gamma P+\frac{\gamma^2}{1-\gamma}R = 33.3 \tag{87}$$

$$q_5 = q_8 = T+\gamma P+\frac{\gamma^2}{1-\gamma}R = 39.3 \tag{88}$$

$$q_6 = q_7 = P+\frac{\gamma}{1-\gamma}R = 37. \tag{89}$$

We can expect that the numerical results converge to the theoretical value in the limit $t\to\infty$. We emphasize that the learned strategy by player 1 is also WSLS, which is consistent with the result in the previous section. We remark that, as the learning proceeds, cooperation by player 1 after the state (C,D) becomes difficult to occur, which leads to the slow convergence of $Q_1(C,C,D)$. On the bottom panel of Figure 1, we provide the numerical results for $\gamma = 0.2 < (T-R)/(R-P)=2/3$. The theoretical value of Q_1 is also provided in

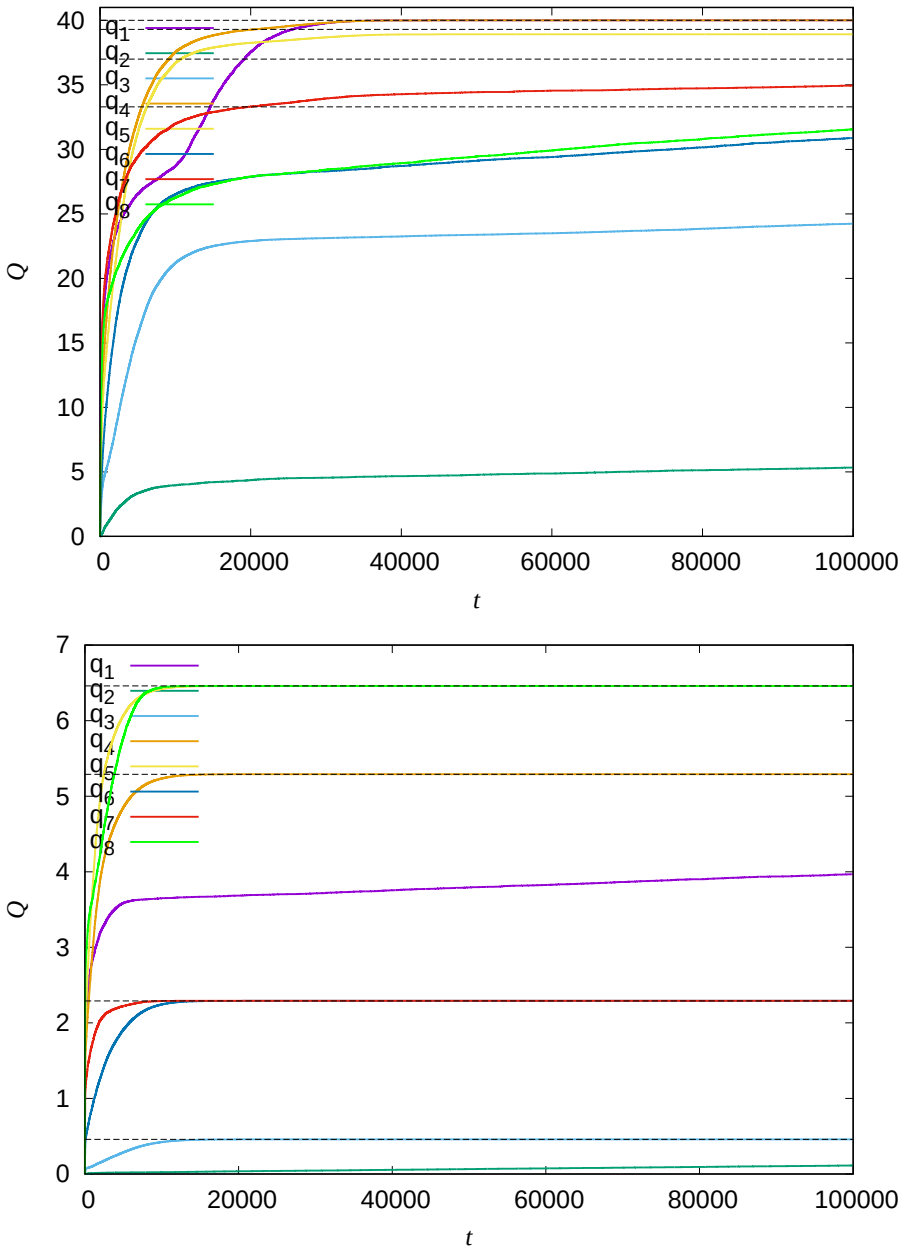

Figure 1: The time evolution of Q_1 when the strategy of player 2 is WSLS. (Top) The value of the discounting factor γ is $\gamma = 0.9$. The straight dash lines correspond to 40, 39.3, 37, and 33.3 from top to bottom. (Bottom) The value of the discounting factor γ is $\gamma = 0.2$. The straight dash lines correspond to 6.46, 5.29, 2.29, and 0.458 from top to bottom.

Appendix A:

$$q_1 = q_4 = R + \frac{\gamma}{1-\gamma^2}T + \frac{\gamma^2}{1-\gamma^2}P \simeq 5.29 \tag{90}$$

$$q_2 = q_3 = S + \frac{\gamma}{1-\gamma^2}P + \frac{\gamma^2}{1-\gamma^2}T \simeq 0.458 \tag{91}$$

$$q_5 = q_8 = \frac{1}{1-\gamma^2}T + \frac{\gamma}{1-\gamma^2}P \simeq 6.46 \tag{92}$$

$$q_6 = q_7 = \frac{1}{1-\gamma^2}P + \frac{\gamma}{1-\gamma^2}T \simeq 2.29. \tag{93}$$

We can expect that the numerical results also converge to the theoretical value in the limit $t \to \infty$. For this case, the learned strategy by player 1 is All-D. Therefore, we conclude that WSLS forms the equilibrium of mutual reinforcement learning for sufficiently large γ.

In Figure 2, we display the time evolution of Q_1 when the strategy of player 2 is Grim. According to Appendix A, the optimal strategy against Grim is Grim for $\gamma > (T-R)/(T-P)$ and All-D for $\gamma < (T-R)/(T-P)$. On the top panel of Figure 2, we provide the numerical results for $\gamma = 0.9 > (T-R)/(T-P) = 2/5$. The theoretical value of Q_1 is also provided in Appendix A:

$$q_1 = \frac{1}{1-\gamma}R = 40 \tag{94}$$

$$q_2 = q_3 = q_4 = S + \frac{\gamma}{1-\gamma}P = 9 \tag{95}$$

$$q_5 = T + \frac{\gamma}{1-\gamma}P = 15 \tag{96}$$

$$q_6 = q_7 = q_8 = \frac{1}{1-\gamma}P = 10. \tag{97}$$

We find that, although the learned strategy by player 1 is Grim, there are discrepancies between the theoretical values and the numerical results for $Q_1(C,C,C)$, $Q_1(C,D,C)$, $Q_1(D,C,C)$, and $Q_1(D,D,C)$. This is due to the property of the Grim strategy. In our simulation, player 1 (a learning agent against Grim) stochastically chooses C or D. However, once player 1 chooses D, player 2 (the agent with the Grim strategy) switches to a defector who always defects. Therefore, the state (D,C) occurs only once. Similarly, the state (C,C) occurs only while player 1 keeps cooperating. Therefore, the number of times that

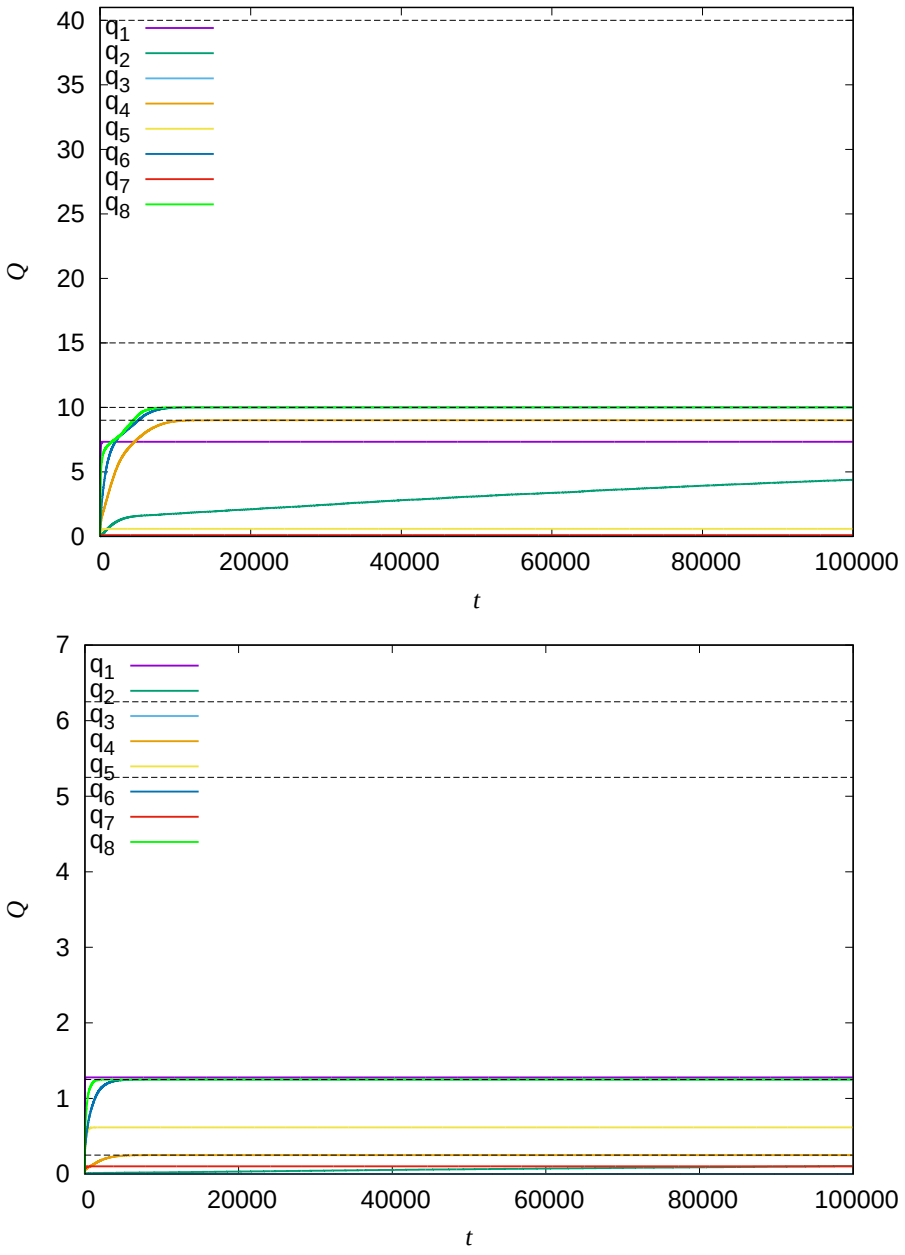

Figure 2: The time evolution of Q_1 when the strategy of player 2 is Grim. (Top) The value of the discounting factor γ is $\gamma = 0.9$. The straight dash lines correspond to 40, 15, 10, and 9 from top to bottom. (Bottom) The value of the discounting factor γ is $\gamma = 0.2$. The straight dash lines correspond to 6.25, 5.25, 1.25, and 0.25 from top to bottom.

the states (C, C) and (D, C) occur in one trial of the infinitely repeated game cannot be large enough for the Q values to converge to the theoretical values. (It has been known that the action-value function in Q-learning converges to the true value if all state-action pairs are visited an infinite number of times [13].) In addition, as the learning proceeds, cooperation by player 1 after the state (C, D) becomes difficult to occur, which leads to the slow convergence of $Q_1(C, C, D)$. On the bottom panel of Figure 2, we provide the numerical results for $\gamma = 0.2 < (T - R)/(T - P) = 2/5$. The theoretical value of Q_1 is also provided in Appendix A:

$$q_1 = R + \gamma T + \frac{\gamma^2}{1-\gamma} P = 5.25 \tag{98}$$

$$q_2 = q_3 = q_4 = S + \frac{\gamma}{1-\gamma} P = 0.25 \tag{99}$$

$$q_5 = T + \frac{\gamma}{1-\gamma} P = 6.25 \tag{100}$$

$$q_6 = q_7 = q_8 = \frac{1}{1-\gamma} P = 1.25. \tag{101}$$

We find that the learned strategy by player 1 is Grim, although the theoretical prediction is All-D. Due to the same reason as above, there are discrepancies between the theoretical values and the numerical results for $Q_1(C, C, C)$, $Q_1(C, D, C)$, $Q_1(D, C, C)$, and $Q_1(D, D, C)$. In particular, although $Q_1(C, C, C)$ is updated as long as player 1 keeps cooperating, $Q_1(D, C, C)$ is updated only once, that is, when player 1 first defects. This fact leads to the discrepancy between the theoretical prediction $Q_1(C, C, C) < Q_1(D, C, C)$ and the numerical result $Q_1(C, C, C) > Q_1(D, C, C)$. (In order to check this conjecture, we also provide numerical results about the situation where implementation error exists in the action of player 2, in Appendix B. These results are consistent with our conjecture.) In addition, due to the same reason as above, the convergence of $Q_1(C, C, D)$ is slow. Besides these facts, our numerical results are consistent with the theoretical prediction, and we conclude that Grim can form the equilibrium of mutual reinforcement learning.

5. Conclusion

In this paper, we theoretically investigated the situation where both players alternately use reinforcement learning to obtain their optimal memory-one strategies in the repeated prisoner's dilemma game. We derived the symmetric solutions of the Bellman optimality equations. We found that WSLS, the Grim strategy, and the All-D strategy can form equilibrium of the mutual reinforcement learning process amongst sixteen deterministic memory-one strategies. We checked this result by numerical simulation using Q-learning. The following problems should be studied in future: (i) Whether asymmetric equilibrium points exist or not, (ii) analysis on non-deterministic strategies, and (iii) extension of our analysis to memory-two strategies. Furthermore, extension of our analysis to the situations where the inequalities $T > R > P > S$ [21, 22, 23, 24] or $2R > T + S$ [25, 26] do not hold is also a subject of future work. In addition, elucidating the relation between equilibrium in the mutual reinforcement learning and equilibrium in evolutionary game theory [15] is a significant problem.

Acknowledgement

This study was supported by JSPS KAKENHI Grant Number JP20K19884.

Appendix A. Optimal strategy against fixed strategies

In this appendix, we provide theoretical results on the deterministic optimal strategy of a learning agent against the other agent with a fixed strategy. We regard agent 1 and 2 as a learning agent and an agent with a fixed strategy, respectively. The Bellman optimality equation of the agent 1 is Eq. (8) as before. We consider the situation where the agent 2 chooses the TFT strategy, the WSLS strategy, and the Grim strategy. We introduce the notation (21) as before.

Appendix A.1. Optimal strategy against TFT

Here we consider the situation that the strategy of the agent 2 is TFT:

$$\boldsymbol{T}_2(C) = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \end{pmatrix}. \tag{A.1}$$

Then, the solution of Eq. (8) is as follows.

Appendix A.1.1. The case $T+S<R+P$ and $\gamma > \frac{P-S}{R-S}$

For the case, the solution is

$$q_1 = q_2 = \frac{1}{1-\gamma}R \tag{A.2}$$

$$q_3 = q_4 = S + \frac{\gamma}{1-\gamma}R \tag{A.3}$$

$$q_5 = q_6 = T + \gamma S + \frac{\gamma^2}{1-\gamma}R \tag{A.4}$$

$$q_7 = q_8 = P + \gamma S + \frac{\gamma^2}{1-\gamma}R \tag{A.5}$$

and because $q_1 > q_5$, $q_2 > q_6$, $q_3 > q_7$, and $q_4 > q_8$, the optimal strategy is All-C.

Appendix A.1.2. The case $T+S<R+P$ and $\frac{T-R}{T-P} < \gamma < \frac{P-S}{R-S}$

For the case, the solution is

$$q_1 = q_2 = \frac{1}{1-\gamma}R \tag{A.6}$$

$$q_3 = q_4 = S + \frac{\gamma}{1-\gamma}R \tag{A.7}$$

$$q_5 = q_6 = T + \frac{\gamma}{1-\gamma}P \tag{A.8}$$

$$q_7 = q_8 = \frac{1}{1-\gamma}P \tag{A.9}$$

and because $q_1 > q_5$, $q_2 > q_6$, $q_3 < q_7$, and $q_4 < q_8$, the optimal strategy is Repeat.

Appendix A.1.3. The case $T+S<R+P$ and $\gamma<\frac{T-R}{T-P}$

For the case, the solution is

$$q_1 = q_2 = R+\gamma T+\frac{\gamma^2}{1-\gamma}P \tag{A.10}$$

$$q_3 = q_4 = S+\gamma T+\frac{\gamma^2}{1-\gamma}P \tag{A.11}$$

$$q_5 = q_6 = T+\frac{\gamma}{1-\gamma}P \tag{A.12}$$

$$q_7 = q_8 = \frac{1}{1-\gamma}P \tag{A.13}$$

and because $q_1 < q_5$, $q_2 < q_6$, $q_3 < q_7$, and $q_4 < q_8$, the optimal strategy is All-D.

Appendix A.1.4. The case $T+S>R+P$ and $\gamma>\frac{T-R}{R-S}$

For the case, the solution is

$$q_1 = q_2 = \frac{1}{1-\gamma}R \tag{A.14}$$

$$q_3 = q_4 = S+\frac{\gamma}{1-\gamma}R \tag{A.15}$$

$$q_5 = q_6 = T+\gamma S+\frac{\gamma^2}{1-\gamma}R \tag{A.16}$$

$$q_7 = q_8 = P+\gamma S+\frac{\gamma^2}{1-\gamma}R \tag{A.17}$$

and because $q_1 > q_5$, $q_2 > q_6$, $q_3 > q_7$, and $q_4 > q_8$, the optimal strategy is All-C.

Appendix A.1.5. The case $T+S>R+P$ and $\frac{P-S}{T-P}<\gamma<\frac{T-R}{R-S}$

For the case, the solution is

$$q_1 = q_2 = R+\frac{\gamma}{1-\gamma^2}T+\frac{\gamma^2}{1-\gamma^2}S \tag{A.18}$$

$$q_3 = q_4 = \frac{1}{1-\gamma^2}S+\frac{\gamma}{1-\gamma^2}T \tag{A.19}$$

$$q_5 = q_6 = \frac{1}{1-\gamma^2}T+\frac{\gamma}{1-\gamma^2}S \tag{A.20}$$

$$q_7 = q_8 = P+\frac{\gamma}{1-\gamma^2}S+\frac{\gamma^2}{1-\gamma^2}T \tag{A.21}$$

and because $q_1 < q_5$, $q_2 < q_6$, $q_3 > q_7$, and $q_4 > q_8$, the optimal strategy is anti-Repeat.

Appendix A.1.6. The case $T+S>R+P$ and $\gamma<\frac{P-S}{T-P}$

For the case, the solution is

$$q_1 = q_2 = R+\gamma T+\frac{\gamma^2}{1-\gamma}P \tag{A.22}$$

$$q_3 = q_4 = S+\gamma T+\frac{\gamma^2}{1-\gamma}P \tag{A.23}$$

$$q_5 = q_6 = T+\frac{\gamma}{1-\gamma}P \tag{A.24}$$

$$q_7 = q_8 = \frac{1}{1-\gamma}P \tag{A.25}$$

and because $q_1 < q_5$, $q_2 < q_6$, $q_3 < q_7$, and $q_4 < q_8$, the optimal strategy is All-D.

Appendix A.2. Optimal strategy against WSLS

Here we consider the situation that the strategy of the agent 2 is WSLS:

$$\boldsymbol{T}_2(C) = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \end{pmatrix}. \tag{A.26}$$

Then, the solution of Eq. (8) is as follows.

Appendix A.2.1. The case $T+P<2R$ and $\gamma>\frac{T-R}{R-P}$

For the case, the solution is

$$q_1 = q_4 = \frac{1}{1-\gamma}R \tag{A.27}$$

$$q_2 = q_3 = S+\gamma P+\frac{\gamma^2}{1-\gamma}R \tag{A.28}$$

$$q_5 = q_8 = T+\gamma P+\frac{\gamma^2}{1-\gamma}R \tag{A.29}$$

$$q_6 = q_7 = P+\frac{\gamma}{1-\gamma}R \tag{A.30}$$

and because $q_1 > q_5$, $q_2 < q_6$, $q_3 < q_7$, and $q_4 > q_8$, the optimal strategy is WSLS.

Appendix A.2.2. The case $T+P<2R$ and $\gamma<\frac{T-R}{R-P}$

For the case, the solution is

$$q_1 = q_4 = R+\frac{\gamma}{1-\gamma^2}T+\frac{\gamma^2}{1-\gamma^2}P \tag{A.31}$$

$$q_2 = q_3 = S+\frac{\gamma}{1-\gamma^2}P+\frac{\gamma^2}{1-\gamma^2}T \tag{A.32}$$

$$q_5 = q_8 = \frac{1}{1-\gamma^2}T+\frac{\gamma}{1-\gamma^2}P \tag{A.33}$$

$$q_6 = q_7 = \frac{1}{1-\gamma^2}P+\frac{\gamma}{1-\gamma^2}T \tag{A.34}$$

and because $q_1<q_5$, $q_2<q_6$, $q_3<q_7$, and $q_4<q_8$, the optimal strategy is All-D.

Appendix A.2.3. The case $T+P>2R$

For the case, the solution is

$$q_1 = q_4 = R+\frac{\gamma}{1-\gamma^2}T+\frac{\gamma^2}{1-\gamma^2}P \tag{A.35}$$

$$q_2 = q_3 = S+\frac{\gamma}{1-\gamma^2}P+\frac{\gamma^2}{1-\gamma^2}T \tag{A.36}$$

$$q_5 = q_8 = \frac{1}{1-\gamma^2}T+\frac{\gamma}{1-\gamma^2}P \tag{A.37}$$

$$q_6 = q_7 = \frac{1}{1-\gamma^2}P+\frac{\gamma}{1-\gamma^2}T \tag{A.38}$$

and because $q_1<q_5$, $q_2<q_6$, $q_3<q_7$, and $q_4<q_8$, the optimal strategy is All-D.

Appendix A.3. Optimal strategy against Grim

Here we consider the situation that the strategy of the agent 2 is Grim:

$$\boldsymbol{T}_2(C) = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}. \tag{A.39}$$

Then, the solution of Eq. (8) is as follows.

Appendix A.3.1. The case $\gamma > \frac{T-R}{T-P}$

For the case, the solution is

$$q_1 = \frac{1}{1-\gamma}R \tag{A.40}$$

$$q_2 = q_3 = q_4 = S + \frac{\gamma}{1-\gamma}P \tag{A.41}$$

$$q_5 = T + \frac{\gamma}{1-\gamma}P \tag{A.42}$$

$$q_6 = q_7 = q_8 = \frac{1}{1-\gamma}P \tag{A.43}$$

and because $q_1 > q_5$, $q_2 < q_6$, $q_3 < q_7$, and $q_4 < q_8$, the optimal strategy is Grim.

Appendix A.3.2. The case $\gamma < \frac{T-R}{T-P}$

For the case, the solution is

$$q_1 = R + \gamma T + \frac{\gamma^2}{1-\gamma}P \tag{A.44}$$

$$q_2 = q_3 = q_4 = S + \frac{\gamma}{1-\gamma}P \tag{A.45}$$

$$q_5 = T + \frac{\gamma}{1-\gamma}P \tag{A.46}$$

$$q_6 = q_7 = q_8 = \frac{1}{1-\gamma}P \tag{A.47}$$

and because $q_1 < q_5$, $q_2 < q_6$, $q_3 < q_7$, and $q_4 < q_8$, the optimal strategy is All-D.

Appendix B. Numerical results under implementation error

In this appendix, we provide numerical results about the situation where implementation error exists in the action of a player. The setup of numerical simulation is the same as one in Section 4. We assume that the strategy of player 2 is Grim with implementation error, which takes wrong action with small probability 10^{-2}. The strategy of player 1 is learned by Q-learning.

In Figure B.3, we display the time evolution of Q_1 when the strategy of player 2 is Grim with implementation error. We can see that the learned strategy of

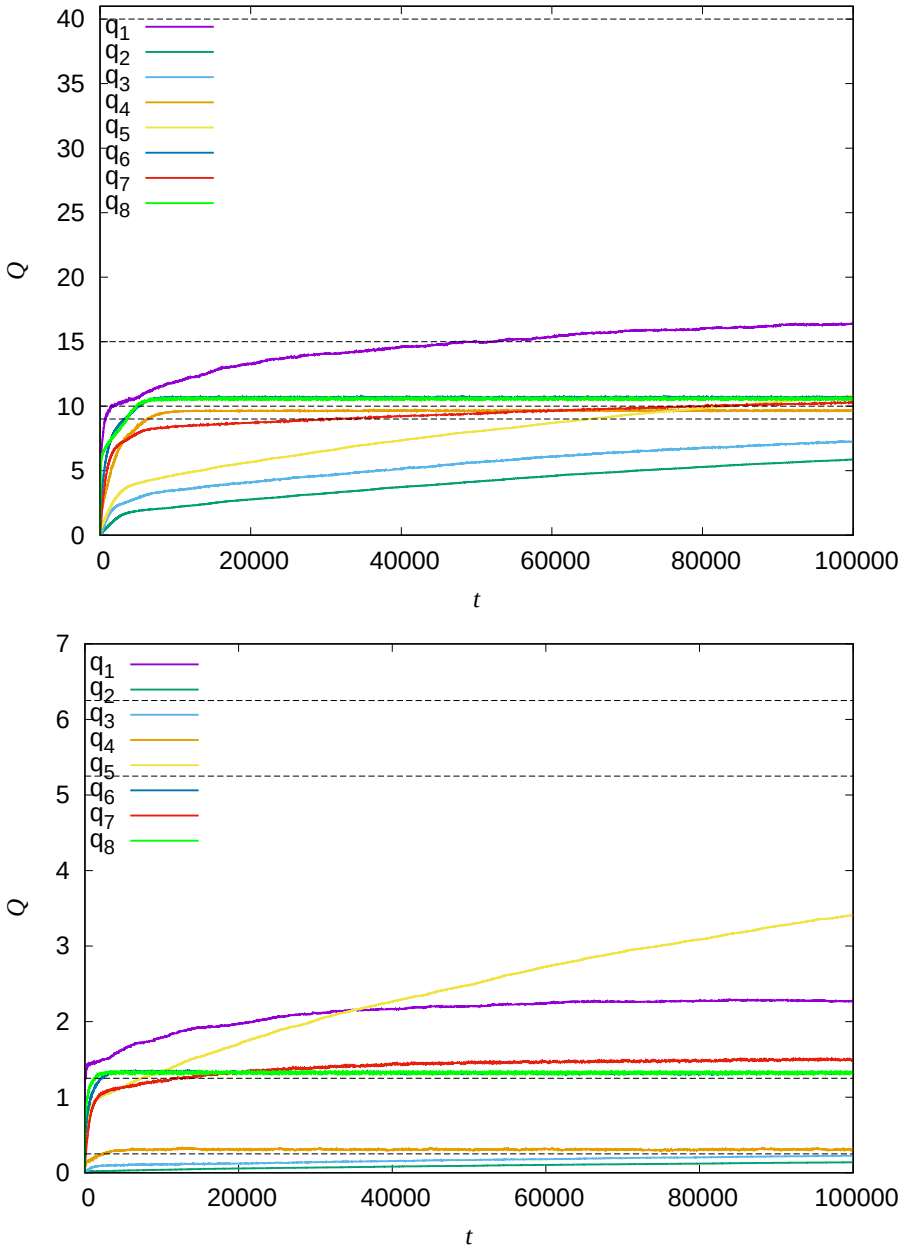

Figure B.3: The time evolution of Q_1 when the strategy of player 2 is Grim with implementation error. (Top) The value of the discounting factor γ is $\gamma = 0.9$. The straight dash lines correspond to 40, 15, 10, and 9 from top to bottom. (Bottom) The value of the discounting factor γ is $\gamma = 0.2$. The straight dash lines correspond to 6.25, 5.25, 1.25, and 0.25 from top to bottom.

player 1 is Grim for $\gamma = 0.9$, and All-D for $\gamma = 0.2$, in contrast to the case without implementation error in Figure 2, where the learned strategy is Grim for both $\gamma = 0.9$ and $\gamma = 0.2$. This is because Grim with implementation error is not irreversible, although Grim is irreversible, and $Q_1(C, C, C)$ and $Q_1(D, C, C)$ are updated sufficiently many times.

References

[1] A. Rapoport, A. M. Chammah, C. J. Orwant, Prisoner's dilemma: A study in conflict and cooperation, Vol. 165, University of Michigan press, 1965.

[2] C. Hilbe, K. Chatterjee, M. A. Nowak, Partners and rivals in direct reciprocity, Nature Human Behaviour 2 (7) (2018) 469.

[3] A. Rapoport, Optimal policies for the prisoner's dilemma., Psychological Review 74 (2) (1967) 136.

[4] T. W. Sandholm, R. H. Crites, Multiagent reinforcement learning in the iterated prisoner's dilemma, Biosystems 37 (1-2) (1996) 147–166.

[5] Y. Sato, E. Akiyama, J. D. Farmer, Chaos in learning a simple two-person game, Proceedings of the National Academy of Sciences 99 (7) (2002) 4748–4751.

[6] J. Hu, M. P. Wellman, Nash q-learning for general-sum stochastic games, Journal of Machine Learning Research 4 (Nov) (2003) 1039–1069.

[7] T. Galla, J. D. Farmer, Complex dynamics in learning complicated games, Proceedings of the National Academy of Sciences 110 (4) (2013) 1232–1236.

[8] S. Hidaka, T. Torii, A. Masumi, Which types of learning make a simple game complex?, Complex Systems 24 (1) (2015) 49–74.

[9] M. Harper, V. Knight, M. Jones, G. Koutsovoulos, N. E. Glynatsi, O. Campbell, Reinforcement learning produces dominant strategies for the iterated prisoner's dilemma, PloS One 12 (12) (2017) e0188046.

[10] W. Barfuss, J. F. Donges, J. Kurths, Deterministic limit of temporal difference reinforcement learning for stochastic games, Physical Review E 99 (4) (2019) 043305.

[11] Y. Fujimoto, K. Kaneko, Emergence of exploitation as symmetry breaking in iterated prisoner's dilemma, Physical Review Research 1 (3) (2019) 033077.

[12] G. I. Bischi, A. Naimzada, Global analysis of a dynamic duopoly game with bounded rationality, in: Advances in dynamic games and applications, Springer, 2000, pp. 361–385.

[13] R. S. Sutton, A. G. Barto, Reinforcement learning: An introduction, MIT press, 2018.

[14] L. Busoniu, R. Babuska, B. De Schutter, A comprehensive survey of multiagent reinforcement learning, IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 38 (2) (2008) 156–172.

[15] J. M. Smith, G. R. Price, The logic of animal conflict, Nature 246 (5427) (1973) 15.

[16] M. Nowak, K. Sigmund, A strategy of win-stay, lose-shift that outperforms tit-for-tat in the prisoner's dilemma game, Nature 364 (6432) (1993) 56–58.

[17] S. Lan, Geometrical regret matching: A new dynamics to nash equilibrium, AIP Advances 10 (6) (2020) 065033.

[18] E. Akin, The iterated prisoner's dilemma: good strategies and their dynamics, Ergodic Theory, Advances in Dynamical Systems (2016) 77–107.

[19] R. Axelrod, W. D. Hamilton, The evolution of cooperation, Science 211 (4489) (1981) 1390–1396.

[20] L. A. Imhof, D. Fudenberg, M. A. Nowak, Tit-for-tat or win-stay, lose-shift?, Journal of Theoretical Biology 247 (3) (2007) 574–580.

[21] J. Tanimoto, H. Sagara, Relationship between dilemma occurrence and the existence of a weakly dominant strategy in a two-player symmetric game, BioSystems 90 (1) (2007) 105–114.

[22] Z. Wang, S. Kokubo, M. Jusup, J. Tanimoto, Universal scaling for the dilemma strength in evolutionary games, Physics of Life Reviews 14 (2015) 1–30.

[23] H. Ito, J. Tanimoto, Scaling the phase-planes of social dilemma strengths shows game-class changes in the five rules governing the evolution of cooperation, Royal Society Open Science 5 (10) (2018) 181085.

[24] M. R. Arefin, K. A. Kabir, M. Jusup, H. Ito, J. Tanimoto, Social efficiency deficit deciphers social dilemmas, Scientific Reports 10 (1) (2020) 1–9.

[25] J. Tanimoto, H. Sagara, A study on emergence of alternating reciprocity in a 2× 2 game with 2-length memory strategy, BioSystems 90 (3) (2007) 728–737.

[26] M. Wakiyama, J. Tanimoto, Reciprocity phase in various 2× 2 games by agents equipped with two-memory length strategy encouraged by grouping for interaction and adaptation, BioSystems 103 (1) (2011) 93–104.