Memory-two strategies forming symmetric mutual reinforcement learning equilibrium in repeated prisoners' dilemma game

Masahiko Ueda

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

Abstract

We investigate symmetric equilibria of mutual reinforcement learning when both players alternately learn the optimal memory-two strategies against the opponent in the repeated prisoners' dilemma game. We provide a necessary condition for memory-two deterministic strategies to form symmetric equilibria. We then provide three examples of memory-two deterministic strategies which form symmetric mutual reinforcement learning equilibria. We also prove that mutual reinforcement learning equilibria formed by memory-two strategies are also mutual reinforcement learning equilibria when both players use reinforcement learning of memory-n strategies with $n > 2$.

Keywords: Repeated prisoners' dilemma game; Reinforcement learning; Memory-two strategies

1. Introduction

The prisoners' dilemma is one of the simplest situations in which rational actions of individuals do not maximize social welfare [1]. Although the best action of each agent is defection, mutual cooperation improves the utility of both agents. On the other hand, if the prisoners' dilemma game is infinitely repeated, the situation changes. In fact, mutual cooperation can be realized by rational behavior of each agent, and this result is known as folk theorem [2]. The folk theorem was also extended to a stronger version that any individually rational payoffs can be realized as subgame perfect equilibria [3].

At the same time, it has been pointed out by experiments that the realistic agents like human beings are not necessarily rational, and theories of bounded rationality have been needed [4]. One of the mainstream is modeling agents by finite automata (agents with finite complexity) [5, 6, 7, 8, 9, 10, 11]. Especially, Abreu and Rubinstein found that the equilibrium payoffs realized by

Email address: m.ueda@yamaguchi-u.ac.jp (Masahiko Ueda)

finite automaton selection games, where players choose finite automata as their strategies in repeated games so as to maximize their payoffs and to minimize the number of states of the finite automata lexicographically, are restricted to some small region in individually rational payoffs [8]. Kalai and Stanford proved that every subgame perfect equilibrium of repeated games can be approximated by a subgame perfect ϵ-equilibrium of finite complexity [7]. A slightly different approach from finite automata is modeling agents by ones with finite memory, which recall only a finite number past periods [12]. (Although there is distinction between memory and recall in computer science, we use these two words interchangeably.) Deterministic finite-memory strategies are contained in a class of finite automata. Sabourian and co-workers investigated how the folk theorem can be extended to finite-memory strategies [13, 14, 15].

Another trend of studies of bounded rationality is modeling agents as adaptive ones which gradually acquire favorable strategies. One of the most successful approach is evolutionary game theory, where a population of individuals evolves by natural selection [16]. The concept of evolutionarily stable strategy, which is interpreted as stability against mutation, succeeded in strengthening the concept of Nash equilibrium. However, it was also shown that any strategy in the infinitely repeated prisoners' dilemma game is not an evolutionarily stable strategy, and is not stable against neutral drift [17]. There are also studies of evolutionarily stable strategies with finite complexity [18, 19, 20]. Particularly, Binmore and Samuelson proposed a modified version of evolutionarily stable strategy and showed that such strategies must maximize the sum of payoffs of two players [19]. Furthermore, many evolutionary simulations on finite-memory strategies have been done for various population sizes, mutation rates, and types of interaction [21, 22, 23, 24, 25, 26]. Stewart and Plotkin proposed the concept to evolutionary robust strategies, which is an extension of evolutionarily stable strategies to systems of finite population size and cannot be selectively replaced by any mutant strategies [27].

Learning is another way of adaptation of human beings, and has also attracted much attention in theoretical economics [28, 29, 30], computer science [31], and complex systems theory [32, 33, 34, 35, 36, 37]. Many methods of learning have been proposed in game theory [38], and compared with experimental results [39, 40, 41]. One of the most popular learning methods is reinforcement learning [42]. In reinforcement learning, an agent gradually learns the optimal policy against a stationary environment. Mutual reinforcement learning in game theory is a more difficult problem since the existence of multiple agents makes an environment nonstationary [43, 44, 45, 46, 47]. Several methods have been proposed for reinforcement learning with multiple agents [48].

Recently, memory-n strategies (n periods memory strategies) with $n > 1$ attract much attention in computational evolutionary game theory, because longer memory enables more complicated behavior [49, 50, 51, 52, 53, 54]. Especially, longer memory enables us to design robust strategies against implementation errors. Since agents in evolutionary biology are organisms, which are far from rational, it has been traditionally assumed that the length of memory of such agents is assumed to be short. This is in contrast to chronology of game theory

in economics, where behaviors of rational and forward-looking agents were first studied and then memory length becomes shorter in order to describe agents with bounded rationality. Because rationality of realistic agents is bounded, shorter-memory strategies will be preferred if complexity is also considered.

Here, we investigate mutual reinforcement learning in the repeated prisoners' dilemma game [1]. More explicitly, we investigate properties of equilibria formed by learning agents when the two agents alternately learn their optimal strategies against the opponent. In the previous study [55], it was found that, among all deterministic memory-one strategies, only the Grim trigger strategy, the Win-Stay Lose-Shift strategy, and the All-D strategy can form symmetric equilibrium of mutual reinforcement learning. A natural question is "How does the set of such equilibria grow as the length of memory increases?". Such direction of research can be useful when we construct strong strategies based on memory-one strategies, as in computational evolutionary game theory. Furthermore, we want to understand mutual reinforcement learning equilibria in terms of strategies, not equilibrium payoffs. However, even whether the above equilibria formed by memory-one strategies are still equilibria in memory-n settings or not has not been known.

In this paper, we extend the analysis of Ref. [55] to memory-two strategies. First, we provide a necessary condition for memory-two deterministic strategies to form symmetric equilibria. Then we provide three non-trivial examples of memory-two deterministic strategies which form symmetric mutual reinforcement learning equilibria. Furthermore, we also prove that mutual reinforcement learning equilibria formed by memory-n' strategies are also mutual reinforcement learning equilibria when both players use reinforcement learning of memory-n strategies with $n > n'$.

This paper is organized as follows. In Section 2, we introduce the repeated prisoners' dilemma game with memory-n strategies, and players using reinforcement learning. In Section 3, we show that the structure of the optimal strategies is constrained by the Bellman optimality equation. In Section 4, we introduce the concepts of mutual reinforcement learning equilibrium and symmetric equilibrium. We then provide a necessary condition for memory-two deterministic strategies to form symmetric equilibria. In Section 5, we provide three examples of memory-two deterministic strategies which form symmetric mutual reinforcement learning equilibria. In Section 6, we show that mutual reinforcement learning equilibria formed by memory-n' strategies are also mutual reinforcement learning equilibria when both players use reinforcement learning of memory-n strategies with $n > n'$. Section 7 is devoted to conclusion.

2. Model

We introduce the repeated prisoners' dilemma game [43]. There are two players (1 and 2) in the game. Each player chooses cooperation (C) or defection (D) on every round. The action of player a is written as $\sigma_a \in \{C, D\}$. We collectively write $\boldsymbol{\sigma} := (\sigma_1, \sigma_2)$, and call $\boldsymbol{\sigma}$ an action profile. We also write the space of all possible action profiles as $\Omega := \{C, D\}^2$. The payoff of player

$a \in \{1,2\}$ when the action profile is $\boldsymbol{\sigma}$ is described as $r_a(\boldsymbol{\sigma})$. The payoffs in the prisoners' dilemma game are given by

$$\begin{aligned}(r_1(C,C), r_1(C,D), r_1(D,C), r_1(D,D)) &= (R,S,T,P) \quad (1)\\ (r_2(C,C), r_2(C,D), r_2(D,C), r_2(D,D)) &= (R,T,S,P) \quad (2)\end{aligned}$$

with $T > R > P > S$ and $2R > T + S$. The (time-independent) memory-n strategy ($n \geq 1$) of player a is described as the conditional probability $T_a\left(\sigma_a | \left[\boldsymbol{\sigma}^{(-m)}\right]_{m=1}^{n}\right)$ of taking action σ_a when the action profiles in the previous n rounds are $\left[\boldsymbol{\sigma}^{(-m)}\right]_{m=1}^{n}$, together with an initial condition, where we have introduced the notation $\left[\boldsymbol{\sigma}^{(-m)}\right]_{m=1}^{n} := \left(\boldsymbol{\sigma}^{(-1)}, \cdots, \boldsymbol{\sigma}^{(-n)}\right)$ from newest to oldest [54]. (As a strategy of bounded rational players, we use finite-memory strategies, not finite automata, because the former allows strategies to be stochastic. Although stochastic strategies are allowed in our framework, we investigate only deterministic strategies in this paper.) We write the length of memory of player a as n_a and define $n := \max\{n_1, n_2\}$. In this paper, we assume that n is finite.

Assumption 1. *Both players use time-independent finite-memory strategies.*

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

We consider the situation that both players learn their optimal strategies against the strategy of the opponent by reinforcement learning [42]. In reinforcement learning, each player learns mapping (called policy) from the action profiles $\left[\boldsymbol{\sigma}^{(-m)}\right]_{m=1}^{n}$ in the previous n rounds to his/her action σ so as to maximize his/her expected future reward. We write the action of player a at round t as $\sigma_a(t)$. In addition, we write $r_a(t) := r_a(\boldsymbol{\sigma}(t))$. We define the action-value function of player a as

$$Q_a\left(\sigma_a, \left[\boldsymbol{\sigma}^{(-m)}\right]_{m=1}^{n}\right) := \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k r_a(t+k) \middle| \sigma_a(t) = \sigma_a, [\boldsymbol{\sigma}(s)]_{s=t-1}^{t-n} = \left[\boldsymbol{\sigma}^{(-m)}\right]_{m=1}^{n}\right], \quad (3)$$

where γ is a discounting factor satisfying $0 \leq \gamma < 1$. The action-value function $Q_a\left(\sigma_a, \left[\boldsymbol{\sigma}^{(-m)}\right]_{m=1}^{n}\right)$ represents the expected future payoffs $\sum_{k=0}^{\infty} \gamma^k r_a(t+k)$ of player a after round t by taking action σ_a when action profiles in the previous n rounds are $\left[\boldsymbol{\sigma}^{(-m)}\right]_{m=1}^{n}$. Therefore, the action-value function suggests the best action in each action profile. It should be noted that the right-hand side does not depend on t. Due to the property of memory-n strategies, the action-value function Q_a obeys the Bellman equation against a fixed strategy T_{-a} of the

opponent:

$$\begin{aligned} & Q_a\left(\sigma_a, \left[\boldsymbol{\sigma}^{(-m)}\right]_{m=1}^{n}\right) \\ = & \sum_{\sigma_{-a}} r_a\left(\boldsymbol{\sigma}\right) T_{-a}\left(\sigma_{-a}|\left[\boldsymbol{\sigma}^{(-m)}\right]_{m=1}^{n}\right) \\ & +\gamma \sum_{\sigma_a'} \sum_{\sigma_{-a}} T_a\left(\sigma_a'|\,\boldsymbol{\sigma}, \left[\boldsymbol{\sigma}^{(-m)}\right]_{m=1}^{n-1}\right) T_{-a}\left(\sigma_{-a}|\left[\boldsymbol{\sigma}^{(-m)}\right]_{m=1}^{n}\right) Q_a\left(\sigma_a', \boldsymbol{\sigma}, \left[\boldsymbol{\sigma}^{(-m)}\right]_{m=1}^{n-1}\right). \end{aligned} \tag{4}$$

See Appendix A for the derivation of Eq. (4). It has been known that the optimal policy T_a^* and the optimal action-value function Q_a^* obeys the following Bellman optimality equation:

$$\begin{aligned} & Q_a^*\left(\sigma_a, \left[\boldsymbol{\sigma}^{(-m)}\right]_{m=1}^{n}\right) \\ = & \sum_{\sigma_{-a}} r_a\left(\boldsymbol{\sigma}\right) T_{-a}\left(\sigma_{-a}|\left[\boldsymbol{\sigma}^{(-m)}\right]_{m=1}^{n}\right) \\ & +\gamma \sum_{\sigma_{-a}} T_{-a}\left(\sigma_{-a}|\left[\boldsymbol{\sigma}^{(-m)}\right]_{m=1}^{n}\right) \max_{\hat{\sigma}} Q_a^*\left(\hat{\sigma}, \boldsymbol{\sigma}, \left[\boldsymbol{\sigma}^{(-m)}\right]_{m=1}^{n-1}\right), \end{aligned} \tag{5}$$

with the support

$$\mathrm{supp} T_a^*\left(\cdot|\left[\boldsymbol{\sigma}^{(-m)}\right]_{m=1}^{n}\right) = \arg\max_{\sigma} Q_a^*\left(\sigma, \left[\boldsymbol{\sigma}^{(-m)}\right]_{m=1}^{n}\right). \tag{6}$$

See Appendix B for the derivation of Eqs. (5) and (6). In other words, in the optimal policy against T_{-a}, player a takes the action σ_a which maximizes the value of $Q_a^*\left(\cdot, \left[\boldsymbol{\sigma}^{(-m)}\right]_{m=1}^{n}\right)$ when the action profiles at the previous n rounds are $\left[\boldsymbol{\sigma}^{(-m)}\right]_{m=1}^{n}$. In Q-learning, which is one of the simplest algorithms of reinforcement learning, it is known that values of action-value functions converge to the solutions of the Bellman optimality equation if all state-action pairs are visited an infinite number of times [42].

We investigate the situation that players infinitely repeat the infinitely-repeated games and players alternately learn their optimal strategies in each game, as in Ref. [55]. We write the optimal strategy and the corresponding optimal action-value function of player a at d-th game as $T_a^{*(d)}$ and $Q_a^{*(d)}$, respectively. Given an initial strategy $T_2^{*(0)}$ of player 2, in the $(2l-1)$-th game ($l \in \mathbb{N}$), player 1 learns $T_1^{*(2l-1)}$ against $T_2^{*(2l-2)}$ by calculating $Q_1^{*(2l-1)}$. In the $2l$-th game, player 2 learns $T_2^{*(2l)}$ against $T_1^{*(2l-1)}$ by calculating $Q_2^{*(2l)}$. We are interested in the fixed points of the dynamics, that is, $T_a^{*(\infty)}$ and $Q_a^{*(\infty)}$.

In this paper, we mainly investigate situations that the support (6) contains only one action, that is, strategies are deterministic. The number of deterministic memory-n strategies in the repeated prisoners' dilemma game is $2^{2^{2n}}$, which increases rapidly as n increases.

3. Structure of optimal strategies

Below we consider only the case $n = 2$. The Bellman optimality equation (5) for $n = 2$ is

$$\begin{aligned} & Q_a^* \left(\sigma_a, \boldsymbol{\sigma}^{(-1)}, \boldsymbol{\sigma}^{(-2)} \right) \\ = & \sum_{\sigma_{-a}} r_a \left(\boldsymbol{\sigma} \right) T_{-a} \left(\sigma_{-a} | \, \boldsymbol{\sigma}^{(-1)}, \boldsymbol{\sigma}^{(-2)} \right) \\ & + \gamma \sum_{\sigma_{-a}} T_{-a} \left(\sigma_{-a} | \, \boldsymbol{\sigma}^{(-1)}, \boldsymbol{\sigma}^{(-2)} \right) \max_{\hat{\sigma}} Q_a^* \left(\hat{\sigma}, \boldsymbol{\sigma}, \boldsymbol{\sigma}^{(-1)} \right) \end{aligned} \tag{7}$$

with

$$\mathrm{supp} T_a^* \left(\cdot | \, \boldsymbol{\sigma}^{(-1)}, \boldsymbol{\sigma}^{(-2)} \right) = \arg\max_{\sigma} Q_a^* \left(\sigma, \boldsymbol{\sigma}^{(-1)}, \boldsymbol{\sigma}^{(-2)} \right). \tag{8}$$

The number of memory-two deterministic strategies is 2^{16}, which is quite large, and therefore we cannot investigate all memory-two deterministic strategies as in the case of memory-one deterministic strategies [55]. Instead, we first investigate general properties of optimal strategies.

We introduce the matrix representation of a strategy:

$$\begin{aligned} & T_a \left(\sigma \right) \\ := & \begin{pmatrix} T_a \left(\sigma | \, (C,C),(C,C) \right) & T_a \left(\sigma | \, (C,C),(C,D) \right) & T_a \left(\sigma | \, (C,C),(D,C) \right) & T_a \left(\sigma | \, (C,C),(D,D) \right) \\ T_a \left(\sigma | \, (C,D),(C,C) \right) & T_a \left(\sigma | \, (C,D),(C,D) \right) & T_a \left(\sigma | \, (C,D),(D,C) \right) & T_a \left(\sigma | \, (C,D),(D,D) \right) \\ T_a \left(\sigma | \, (D,C),(C,C) \right) & T_a \left(\sigma | \, (D,C),(C,D) \right) & T_a \left(\sigma | \, (D,C),(D,C) \right) & T_a \left(\sigma | \, (D,C),(D,D) \right) \\ T_a \left(\sigma | \, (D,D),(C,C) \right) & T_a \left(\sigma | \, (D,D),(C,D) \right) & T_a \left(\sigma | \, (D,D),(D,C) \right) & T_a \left(\sigma | \, (D,D),(D,D) \right) \end{pmatrix}. \end{aligned} \tag{9}$$

For deterministic strategies, each component in the matrix is 0 or 1.

We now prove the following proposition:

Proposition 1. *For two different action profiles $\boldsymbol{\sigma}^{(-2)}$ and $\boldsymbol{\sigma}^{(-2)\prime}$, if*

$$T_{-a} \left(\sigma | \, \boldsymbol{\sigma}^{(-1)}, \boldsymbol{\sigma}^{(-2)} \right) = T_{-a} \left(\sigma | \, \boldsymbol{\sigma}^{(-1)}, \boldsymbol{\sigma}^{(-2)\prime} \right) \quad (\forall \sigma) \tag{10}$$

holds for some $\boldsymbol{\sigma}^{(-1)}$, then

$$T_a^* \left(\sigma | \, \boldsymbol{\sigma}^{(-1)}, \boldsymbol{\sigma}^{(-2)} \right) = T_a^* \left(\sigma | \, \boldsymbol{\sigma}^{(-1)}, \boldsymbol{\sigma}^{(-2)\prime} \right) \quad (\forall \sigma) \tag{11}$$

also holds.

Proof. For such $\boldsymbol{\sigma}^{(-1)}$, because of Eq. (7), we find

$$\begin{aligned}
& Q_a^*\left(\sigma_a, \boldsymbol{\sigma}^{(-1)}, \boldsymbol{\sigma}^{(-2)}\right) \\
= & \sum_{\sigma_{-a}} r_a\left(\boldsymbol{\sigma}\right) T_{-a}\left(\sigma_{-a} | \, \boldsymbol{\sigma}^{(-1)}, \boldsymbol{\sigma}^{(-2)}\right) \\
& +\gamma \sum_{\sigma_{-a}} T_{-a}\left(\sigma_{-a} | \, \boldsymbol{\sigma}^{(-1)}, \boldsymbol{\sigma}^{(-2)}\right) \max_{\hat{\sigma}} Q_a^*\left(\hat{\sigma}, \boldsymbol{\sigma}, \boldsymbol{\sigma}^{(-1)}\right) \\
= & \sum_{\sigma_{-a}} r_a\left(\boldsymbol{\sigma}\right) T_{-a}\left(\sigma_{-a} | \, \boldsymbol{\sigma}^{(-1)}, \boldsymbol{\sigma}^{(-2)\prime}\right) \\
& +\gamma \sum_{\sigma_{-a}} T_{-a}\left(\sigma_{-a} | \, \boldsymbol{\sigma}^{(-1)}, \boldsymbol{\sigma}^{(-2)\prime}\right) \max_{\hat{\sigma}} Q_a^*\left(\hat{\sigma}, \boldsymbol{\sigma}, \boldsymbol{\sigma}^{(-1)}\right) \\
= & Q_a^*\left(\sigma_a, \boldsymbol{\sigma}^{(-1)}, \boldsymbol{\sigma}^{(-2)\prime}\right) \qquad (12)
\end{aligned}$$

for all σ_a. Since T_a^* is determined by Eq. (8), we obtain Eq. (11). □

This proposition implies that the structure of the matrix $T_a^*(\sigma)$ is the same as that of $T_{-a}(\sigma)$. For deterministic strategies, in order to see this in more detail, we introduce the following sets for $a \in \{1,2\}$ and $\boldsymbol{\sigma}^{(-1)} \in \Omega$:

$$N_x^{(a)}\left(\boldsymbol{\sigma}^{(-1)}\right) := \left\{ \boldsymbol{\sigma}^{(-2)} \in \Omega \middle| \, T_a\left(C | \, \boldsymbol{\sigma}^{(-1)}, \boldsymbol{\sigma}^{(-2)}\right) = x \right\}, \qquad (13)$$

where $x \in \{0,1\}$. That is, $N_1^{(a)}\left(\boldsymbol{\sigma}^{(-1)}\right)$ describes the set of $\boldsymbol{\sigma}^{(-2)}$ such that player a using strategy T_a cooperates after the history $\left[\boldsymbol{\sigma}^{(-m)}\right]_{m=1}^{2}$. Similarly, $N_0^{(a)}\left(\boldsymbol{\sigma}^{(-1)}\right)$ describes the set of $\boldsymbol{\sigma}^{(-2)}$ such that player a using strategy T_a defects after the history $\left[\boldsymbol{\sigma}^{(-m)}\right]_{m=1}^{2}$. We remark that $N_0^{(a)}\left(\boldsymbol{\sigma}^{(-1)}\right) \cup N_1^{(a)}\left(\boldsymbol{\sigma}^{(-1)}\right) = \Omega$ for all a and $\boldsymbol{\sigma}^{(-1)}$. Then, Proposition 1 leads the following corollary:

Corollary 1. *For a deterministic strategy T_{-a} of player $-a$, if the optimal strategy T_a^* of player a against T_{-a} is also deterministic, then one of the following four relations holds for each $\boldsymbol{\sigma}^{(-1)} \in \Omega$:*

(a) $N_x^{(a)}\left(\boldsymbol{\sigma}^{(-1)}\right) = N_x^{(-a)}\left(\boldsymbol{\sigma}^{(-1)}\right)$ *for all* x
(b) $N_x^{(a)}\left(\boldsymbol{\sigma}^{(-1)}\right) = N_{1-x}^{(-a)}\left(\boldsymbol{\sigma}^{(-1)}\right)$ *for all* x
(c) $N_0^{(a)}\left(\boldsymbol{\sigma}^{(-1)}\right) = N_0^{(-a)}\left(\boldsymbol{\sigma}^{(-1)}\right) \cup N_1^{(-a)}\left(\boldsymbol{\sigma}^{(-1)}\right) = \Omega$ *and* $N_1^{(a)}\left(\boldsymbol{\sigma}^{(-1)}\right) = \emptyset$
(d) $N_1^{(a)}\left(\boldsymbol{\sigma}^{(-1)}\right) = N_0^{(-a)}\left(\boldsymbol{\sigma}^{(-1)}\right) \cup N_1^{(-a)}\left(\boldsymbol{\sigma}^{(-1)}\right) = \Omega$ *and* $N_0^{(a)}\left(\boldsymbol{\sigma}^{(-1)}\right) = \emptyset$.

4. Symmetric equilibrium

In this section, we investigate symmetric equilibrium of mutual reinforcement learning.

First, we introduce the notation $\overline{C} := D$, $\overline{D} := C$, and $\pi\left(\sigma_1, \sigma_2\right) := \left(\sigma_2, \sigma_1\right)$. We define the word *same* strategy.

Definition 1. *A strategy T_a of player a is the* same *strategy as that of player $-a$ iff*

$$T_a\left(\sigma|\,\boldsymbol{\sigma}^{(-1)},\boldsymbol{\sigma}^{(-2)}\right) \;=\; T_{-a}\left(\sigma|\,\pi\left(\boldsymbol{\sigma}^{(-1)}\right),\pi\left(\boldsymbol{\sigma}^{(-2)}\right)\right) \tag{14}$$

for all σ, $\boldsymbol{\sigma}^{(-1)}$ and, $\boldsymbol{\sigma}^{(-2)}$.

Next, we introduce equilibria achieved by mutual reinforcement learning.

Definition 2. *A pair of strategy T_1 and T_2 is a* mutual reinforcement learning equilibrium *iff T_a is the optimal strategy against T_{-a} for $a=1,2$.*

We emphasize that such equilibria are defined only for a time-independent part of finite-memory strategies T_a, although finite-memory strategies of players are generally defined as a pair of a time-independent part T_a and an initial condition. This definition is in contrast to that of Nash equilibrium or subgame perfect equilibrium. When some appropriate initial condition is chosen, it becomes a subgame perfect equilibrium of all time-independent finite-memory strategies. In addition, because the optimal policy is determined by comparing the action-value functions, which are functions of finite-length histories including off-equilibrium path, mutual reinforcement learning equilibrium is quite different from Nash equilibrium.

We also remark that a mutual reinforcement learning equilibrium can be achieved by Q-learning if all state-action pairs are visited an infinite number of times as mentioned above, and if an initial strategy of player 2 is appropriate. Even if not all state-action pairs are visited an infinite number of times, we can obtain the mutual reinforcement learning equilibrium by introducing infinitesimal error probability to the opponent's strategy as in Ref. [55].

For deterministic mutual reinforcement learning equilibria, the following proposition is the direct consequence of Corollary 1.

Proposition 2. *For mutual reinforcement learning equilibria formed by deterministic strategies, one of the following two relations holds for each $\boldsymbol{\sigma}^{(-1)}\in\Omega$:*

(a) $N_x^{(1)}\left(\boldsymbol{\sigma}^{(-1)}\right)=N_x^{(2)}\left(\boldsymbol{\sigma}^{(-1)}\right)$ *for all x*
(b) $N_x^{(1)}\left(\boldsymbol{\sigma}^{(-1)}\right)=N_{1-x}^{(2)}\left(\boldsymbol{\sigma}^{(-1)}\right)$ *for all x.*

Proof. According to Corollary 1, one of the four situations (a)-(d) holds for the optimal strategy T_1 against T_2. However, because T_2 is also the optimal strategy against T_1, the cases (c) and (d) are excluded or integrated into the case (a) or (b). □

Furthermore, we introduce symmetric equilibria of mutual reinforcement learning.

Definition 3. *A pair of strategy T_1 and T_2 is a* symmetric mutual reinforcement learning equilibrium *iff T_a is the optimal strategy against T_{-a} and T_a is the same strategy as T_{-a} for $a=1,2$.*

It should be noted that the deterministic optimal strategies can be written as

$$T_a^*\left(\sigma|\,\boldsymbol{\sigma}^{(-1)},\boldsymbol{\sigma}^{(-2)}\right) \;=\; \mathbb{I}\left(Q_a^*\left(\sigma,\boldsymbol{\sigma}^{(-1)},\boldsymbol{\sigma}^{(-2)}\right) > Q_a^*\left(\overline{\sigma},\boldsymbol{\sigma}^{(-1)},\boldsymbol{\sigma}^{(-2)}\right)\right), \tag{15}$$

where $\mathbb{I}(\cdots)$ is the indicator function that returns 1 when $\cdots$ holds and 0 otherwise. We also introduce the following sets for $a \in \{1,2\}$ and $\boldsymbol{\sigma}^{(-1)} \in \Omega$:

$$\tilde{N}_x^{(a)}\left(\boldsymbol{\sigma}^{(-1)}\right) \;:=\; \left\{\boldsymbol{\sigma}^{(-2)} \in \Omega \middle| \, T_a\left(C|\,\boldsymbol{\sigma}^{(-1)},\pi\left(\boldsymbol{\sigma}^{(-2)}\right)\right) = x\right\}, \tag{16}$$

where $x \in \{0,1\}$. We now prove the first main result of this paper.

Theorem 1. *For symmetric mutual reinforcement learning equilibria formed by deterministic strategies, the following relations must hold:*

(a) *For* $\boldsymbol{\sigma}^{(-1)} \in \{(C,C),(D,D)\}$,

$$T_a\left(C|\,\boldsymbol{\sigma}^{(-1)},(C,D)\right) \;=\; T_a\left(C|\,\boldsymbol{\sigma}^{(-1)},(D,C)\right) \tag{17}$$

for all a.

(b) *For* $\boldsymbol{\sigma}^{(-1)} \in \{(C,D),(D,C)\}$,

$$N_x^{(a)}\left(\pi\left(\boldsymbol{\sigma}^{(-1)}\right)\right) \;=\; \tilde{N}_x^{(a)}\left(\boldsymbol{\sigma}^{(-1)}\right) \quad (\forall x) \tag{18}$$

or

$$N_x^{(a)}\left(\pi\left(\boldsymbol{\sigma}^{(-1)}\right)\right) \;=\; \tilde{N}_{1-x}^{(a)}\left(\boldsymbol{\sigma}^{(-1)}\right) \quad (\forall x) \tag{19}$$

holds.

Proof. For $\boldsymbol{\sigma}^{(-1)} \in \{(C,C),(D,D)\}$, $\pi\left(\boldsymbol{\sigma}^{(-1)}\right) = \boldsymbol{\sigma}^{(-1)}$ holds. Because T_1 and T_2 are the same strategies as each other,

$$T_1\left(C|\,\boldsymbol{\sigma}^{(-1)},\boldsymbol{\sigma}^{(-2)}\right) \;=\; T_2\left(C|\,\boldsymbol{\sigma}^{(-1)},\boldsymbol{\sigma}^{(-2)}\right) \quad \left(\boldsymbol{\sigma}^{(-2)} \in \{(C,C),(D,D)\}\right) \tag{20}$$

holds. This and Proposition 2 imply that $N_x^{(1)}\left(\boldsymbol{\sigma}^{(-1)}\right) = N_x^{(2)}\left(\boldsymbol{\sigma}^{(-1)}\right)$ $(\forall x \in \{0,1\})$ must holds. On the other hand, due to Eq. (14),

$$T_1\left(C|\,\boldsymbol{\sigma}^{(-1)},(C,D)\right) \;=\; T_2\left(C|\,\boldsymbol{\sigma}^{(-1)},(D,C)\right) \tag{21}$$

$$T_1\left(C|\,\boldsymbol{\sigma}^{(-1)},(D,C)\right) \;=\; T_2\left(C|\,\boldsymbol{\sigma}^{(-1)},(C,D)\right) \tag{22}$$

also hold. This means that, if $(C,D) \in N_x^{(1)}\left(\boldsymbol{\sigma}^{(-1)}\right)$, then $(D,C) \in N_x^{(2)}\left(\pi\left(\boldsymbol{\sigma}^{(-1)}\right)\right) = N_x^{(2)}\left(\boldsymbol{\sigma}^{(-1)}\right) = N_x^{(1)}\left(\boldsymbol{\sigma}^{(-1)}\right)$, leading to Eq. (17).

For $\boldsymbol{\sigma}^{(-1)} \in \{(C,D),(D,C)\}$, because T_1 and T_2 are the same strategies as each other,

$$T_2\left(C|\,\pi\left(\boldsymbol{\sigma}^{(-1)}\right),\boldsymbol{\sigma}^{(-2)}\right) \;=\; T_1\left(C|\,\boldsymbol{\sigma}^{(-1)},\pi\left(\boldsymbol{\sigma}^{(-2)}\right)\right) \tag{23}$$

holds for $\forall\boldsymbol{\sigma}^{(-2)} \in \Omega$. This means that

$$N_x^{(2)}\left(\pi\left(\boldsymbol{\sigma}^{(-1)}\right)\right) \;=\; \tilde{N}_x^{(1)}\left(\boldsymbol{\sigma}^{(-1)}\right) \quad (\forall x \in \{0,1\}) \tag{24}$$

holds. On the other hand, Proposition 2 implies that

$$N_x^{(1)}\left(\pi\left(\boldsymbol{\sigma}^{(-1)}\right)\right) \;=\; N_x^{(2)}\left(\pi\left(\boldsymbol{\sigma}^{(-1)}\right)\right) \quad (\forall x \in \{0,1\}) \tag{25}$$

or

$$N_x^{(1)}\left(\pi\left(\boldsymbol{\sigma}^{(-1)}\right)\right) \;=\; N_{1-x}^{(2)}\left(\pi\left(\boldsymbol{\sigma}^{(-1)}\right)\right) \quad (\forall x \in \{0,1\}) \tag{26}$$

must hold. By combining Eq. (24) and Eq. (25) or (26), we obtain Eq. (18) or (19). □

Theorem 1 provides a necessary condition for a deterministic strategy to form a symmetric mutual reinforcement learning equilibrium. In particular, Eqs. (18) and (19) imply that the second row and the third row of T_a cannot be independent of each other. Explicitly, T_a must be one of the following 8 forms:

$$\begin{pmatrix} a_1 & b_1 & b_1 & a_2 \\ c_1 & c_1 & c_1 & c_1 \\ d_1 & d_1 & d_1 & d_1 \\ a_3 & b_2 & b_2 & a_4 \end{pmatrix}, \quad \begin{pmatrix} a_1 & b_1 & b_1 & a_2 \\ c_1 & c_1 & c_1 & 1-c_1 \\ d_1 & d_1 & d_1 & 1-d_1 \\ a_3 & b_2 & b_2 & a_4 \end{pmatrix},$$
$$\begin{pmatrix} a_1 & b_1 & b_1 & a_2 \\ c_1 & c_1 & 1-c_1 & c_1 \\ d_1 & 1-d_1 & d_1 & d_1 \\ a_3 & b_2 & b_2 & a_4 \end{pmatrix}, \quad \begin{pmatrix} a_1 & b_1 & b_1 & a_2 \\ c_1 & c_1 & 1-c_1 & 1-c_1 \\ d_1 & 1-d_1 & d_1 & 1-d_1 \\ a_3 & b_2 & b_2 & a_4 \end{pmatrix},$$
$$\begin{pmatrix} a_1 & b_1 & b_1 & a_2 \\ c_1 & 1-c_1 & c_1 & c_1 \\ d_1 & d_1 & 1-d_1 & d_1 \\ a_3 & b_2 & b_2 & a_4 \end{pmatrix}, \quad \begin{pmatrix} a_1 & b_1 & b_1 & a_2 \\ c_1 & 1-c_1 & c_1 & 1-c_1 \\ d_1 & d_1 & 1-d_1 & 1-d_1 \\ a_3 & b_2 & b_2 & a_4 \end{pmatrix},$$
$$\begin{pmatrix} a_1 & b_1 & b_1 & a_2 \\ c_1 & 1-c_1 & 1-c_1 & c_1 \\ d_1 & 1-d_1 & 1-d_1 & d_1 \\ a_3 & b_2 & b_2 & a_4 \end{pmatrix}, \quad \begin{pmatrix} a_1 & b_1 & b_1 & a_2 \\ c_1 & 1-c_1 & 1-c_1 & 1-c_1 \\ d_1 & 1-d_1 & 1-d_1 & 1-d_1 \\ a_3 & b_2 & b_2 & a_4 \end{pmatrix}, \tag{27}$$

where $a_i, b_j, c_1, d_1 \in \{0,1\}$ $(i = 1,2,3,4)$, $(j = 1,2)$ independently. For example, the Tit-for-Tat-anti-Tit-for-Tat (TFT-ATFT) strategy [50]

$$T_1(C) \;=\; \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix}, \tag{28}$$

does not satisfy the condition of Theorem 1, and therefore it cannot form a symmetric mutual reinforcement learning equilibrium. However, there are still many memory-two strategies which satisfy the necessary condition, and further refinement will be needed.

5. Examples of deterministic strategies forming symmetric equilibrium

In this section, we provide three examples of memory-two deterministic strategies forming symmetric mutual reinforcement learning equilibrium. For convenience, we define the following sixteen quantities:

$$q_1 := R + \gamma \max_{\sigma} Q_1^* (\sigma, (C,C), (C,C)) \tag{29}$$
$$q_2 := T + \gamma \max_{\sigma} Q_1^* (\sigma, (D,C), (C,C)) \tag{30}$$
$$q_3 := S + \gamma \max_{\sigma} Q_1^* (\sigma, (C,D), (C,C)) \tag{31}$$
$$q_4 := P + \gamma \max_{\sigma} Q_1^* (\sigma, (D,D), (C,C)) \tag{32}$$

$$q_5 := R + \gamma \max_{\sigma} Q_1^* (\sigma, (C,C), (C,D)) \tag{33}$$
$$q_6 := T + \gamma \max_{\sigma} Q_1^* (\sigma, (D,C), (C,D)) \tag{34}$$
$$q_7 := S + \gamma \max_{\sigma} Q_1^* (\sigma, (C,D), (C,D)) \tag{35}$$
$$q_8 := P + \gamma \max_{\sigma} Q_1^* (\sigma, (D,D), (C,D)) \tag{36}$$

$$q_9 := R + \gamma \max_{\sigma} Q_1^* (\sigma, (C,C), (D,C)) \tag{37}$$
$$q_{10} := T + \gamma \max_{\sigma} Q_1^* (\sigma, (D,C), (D,C)) \tag{38}$$
$$q_{11} := S + \gamma \max_{\sigma} Q_1^* (\sigma, (C,D), (D,C)) \tag{39}$$
$$q_{12} := P + \gamma \max_{\sigma} Q_1^* (\sigma, (D,D), (D,C)) \tag{40}$$

$$q_{13} := R + \gamma \max_{\sigma} Q_1^* (\sigma, (C,C), (D,D)) \tag{41}$$
$$q_{14} := T + \gamma \max_{\sigma} Q_1^* (\sigma, (D,C), (D,D)) \tag{42}$$
$$q_{15} := S + \gamma \max_{\sigma} Q_1^* (\sigma, (C,D), (D,D)) \tag{43}$$
$$q_{16} := P + \gamma \max_{\sigma} Q_1^* (\sigma, (D,D), (D,D)) \tag{44}$$

The Bellman optimality equation for symmetric equilibrium is

$$\begin{aligned} & Q_1^*\left(\sigma_1, \boldsymbol{\sigma}^{(-1)}, \boldsymbol{\sigma}^{(-2)}\right) \\ = & \sum_{\sigma_2}\left\{r_1(\boldsymbol{\sigma}) + \max_{\hat{\sigma}} Q_1^*\left(\hat{\sigma}, \boldsymbol{\sigma}, \boldsymbol{\sigma}^{(-1)}\right)\right\} \\ & \times \mathbb{I}\left(Q_1^*\left(\sigma_2, \pi\left(\boldsymbol{\sigma}^{(-1)}\right), \pi\left(\boldsymbol{\sigma}^{(-2)}\right)\right) > Q_1^*\left(\overline{\sigma}_2, \pi\left(\boldsymbol{\sigma}^{(-1)}\right), \pi\left(\boldsymbol{\sigma}^{(-2)}\right)\right)\right). \end{aligned} \tag{45}$$

We want to find solutions of this equation.

5.1. Delayed Grim trigger strategy

The first candidate of the solution of Eq. (45) is

$$T_1(C) = T_2(C) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}. \tag{46}$$

We can easily check that this strategy satisfies the necessary condition for symmetric equilibrium in Theorem 1. Because this strategy is a variant of the Grim trigger strategy [56]

$$T_1(C) = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \tag{47}$$

but uses only information at the second last action profile, the strategy (46) can be called as *delayed Grim* strategy.

Theorem 2. *A pair of the strategy (46) forms a symmetric mutual reinforcement learning equilibrium if* $\gamma > \sqrt{\frac{T-R}{T-P}}$.

Proof. The Bellman optimality equation against the strategy (46) is

$$Q_1^*(C,(C,C),(C,C)) = q_1 \tag{48}$$
$$Q_1^*(D,(C,C),(C,C)) = q_2 \tag{49}$$
$$Q_1^*\left(C,(C,C),\boldsymbol{\sigma}^{(-2)}\right) = q_3 \quad \left(\boldsymbol{\sigma}^{(-2)} \neq (C,C)\right) \tag{50}$$
$$Q_1^*\left(D,(C,C),\boldsymbol{\sigma}^{(-2)}\right) = q_4 \quad \left(\boldsymbol{\sigma}^{(-2)} \neq (C,C)\right) \tag{51}$$

$$Q_1^*(C,(C,D),(C,C)) = q_5 \tag{52}$$
$$Q_1^*(D,(C,D),(C,C)) = q_6 \tag{53}$$
$$Q_1^*\left(C,(C,D),\boldsymbol{\sigma}^{(-2)}\right) = q_7 \quad \left(\boldsymbol{\sigma}^{(-2)} \neq (C,C)\right) \tag{54}$$
$$Q_1^*\left(D,(C,D),\boldsymbol{\sigma}^{(-2)}\right) = q_8 \quad \left(\boldsymbol{\sigma}^{(-2)} \neq (C,C)\right) \tag{55}$$

$$
\begin{aligned}
Q_1^*\left(C,(D,C),(C,C)\right) &= q_9 && (56)\\
Q_1^*\left(D,(D,C),(C,C)\right) &= q_{10} && (57)\\
Q_1^*\left(C,(D,C),\boldsymbol{\sigma}^{(-2)}\right) &= q_{11} \quad \left(\boldsymbol{\sigma}^{(-2)} \neq (C,C)\right) && (58)\\
Q_1^*\left(D,(D,C),\boldsymbol{\sigma}^{(-2)}\right) &= q_{12} \quad \left(\boldsymbol{\sigma}^{(-2)} \neq (C,C)\right) && (59)
\end{aligned}
$$

$$
\begin{aligned}
Q_1^*\left(C,(D,D),(C,C)\right) &= q_{13} && (60)\\
Q_1^*\left(D,(D,D),(C,C)\right) &= q_{14} && (61)\\
Q_1^*\left(C,(D,D),\boldsymbol{\sigma}^{(-2)}\right) &= q_{15} \quad \left(\boldsymbol{\sigma}^{(-2)} \neq (C,C)\right) && (62)\\
Q_1^*\left(D,(D,D),\boldsymbol{\sigma}^{(-2)}\right) &= q_{16} \quad \left(\boldsymbol{\sigma}^{(-2)} \neq (C,C)\right) && (63)
\end{aligned}
$$

with the self-consistency condition

$$
\begin{aligned}
q_1 &> q_2\\
q_3 &< q_4\\
q_5 &> q_6\\
q_7 &< q_8\\
q_9 &> q_{10}\\
q_{11} &< q_{12}\\
q_{13} &> q_{14}\\
q_{15} &< q_{16}. \qquad (64)
\end{aligned}
$$

The solution is

$$
\begin{aligned}
q_1 &= \frac{1}{1-\gamma}R && (65)\\
q_2 &= T + \frac{\gamma}{1-\gamma^2}R + \frac{\gamma^2}{1-\gamma^2}P && (66)\\
q_3 &= S + \frac{\gamma}{1-\gamma^2}R + \frac{\gamma^2}{1-\gamma^2}P && (67)\\
q_4 &= \frac{1}{1-\gamma^2}P + \frac{\gamma}{1-\gamma^2}R && (68)\\
q_5 = q_9 = q_{13} &= \frac{1}{1-\gamma^2}R + \frac{\gamma}{1-\gamma^2}P && (69)\\
q_6 = q_{10} = q_{14} &= T + \frac{\gamma}{1-\gamma}P && (70)\\
q_7 = q_{11} = q_{15} &= S + \frac{\gamma}{1-\gamma}P && (71)\\
q_8 = q_{12} = q_{16} &= \frac{1}{1-\gamma}P. && (72)
\end{aligned}
$$

For these solution, the inequalities (64) are satisfied if

$$\gamma \quad > \quad \sqrt{\frac{T-R}{T-P}}. \tag{73}$$

□

We remark that the condition (73) is more strict than the condition that Grim forms a symmetric equilibrium [55]: $\gamma > \frac{T-R}{T-P}$.

5.2. Delayed Win-Stay Lose-Shift strategy

The second candidate of the solution of Eq. (45) is

$$T_1(C) = T_2(C) \quad = \quad \begin{pmatrix} 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix}. \tag{74}$$

We can easily check that this strategy satisfies the necessary condition for symmetric equilibrium in Theorem 1. Because this strategy is a variant of the Win-Stay Lose-Shift (WSLS) strategy [22]

$$T_1(C) \quad = \quad \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix} \tag{75}$$

but uses only information at the second last action profile, the strategy (74) can be called as *delayed WSLS* strategy.

Theorem 3. *When $2R > T + P$ holds, a pair of the strategy (74) forms a symmetric mutual reinforcement learning equilibrium if $\gamma > \sqrt{\frac{T-R}{R-P}}$.*

Proof. The Bellman optimality equation against the strategy (74) is

$$Q_1^*\left(C,(C,C),\boldsymbol{\sigma}^{(-2)}\right) = q_1 \quad \left(\boldsymbol{\sigma}^{(-2)} \in \{(C,C),(D,D)\}\right) \tag{76}$$

$$Q_1^*\left(D,(C,C),\boldsymbol{\sigma}^{(-2)}\right) = q_2 \quad \left(\boldsymbol{\sigma}^{(-2)} \in \{(C,C),(D,D)\}\right) \tag{77}$$

$$Q_1^*\left(C,(C,C),\boldsymbol{\sigma}^{(-2)}\right) = q_3 \quad \left(\boldsymbol{\sigma}^{(-2)} \in \{(C,D),(D,C)\}\right) \tag{78}$$

$$Q_1^*\left(D,(C,C),\boldsymbol{\sigma}^{(-2)}\right) = q_4 \quad \left(\boldsymbol{\sigma}^{(-2)} \in \{(C,D),(D,C)\}\right) \tag{79}$$

$$Q_1^*\left(C,(C,D),\boldsymbol{\sigma}^{(-2)}\right) = q_5 \quad \left(\boldsymbol{\sigma}^{(-2)} \in \{(C,C),(D,D)\}\right) \tag{80}$$

$$Q_1^*\left(D,(C,D),\boldsymbol{\sigma}^{(-2)}\right) = q_6 \quad \left(\boldsymbol{\sigma}^{(-2)} \in \{(C,C),(D,D)\}\right) \tag{81}$$

$$Q_1^*\left(C,(C,D),\boldsymbol{\sigma}^{(-2)}\right) = q_7 \quad \left(\boldsymbol{\sigma}^{(-2)} \in \{(C,D),(D,C)\}\right) \tag{82}$$

$$Q_1^*\left(D,(C,D),\boldsymbol{\sigma}^{(-2)}\right) = q_8 \quad \left(\boldsymbol{\sigma}^{(-2)} \in \{(C,D),(D,C)\}\right) \tag{83}$$

$$
\begin{aligned}
Q_1^*\left(C,(D,C),\boldsymbol{\sigma}^{(-2)}\right) &= q_9 \quad \left(\boldsymbol{\sigma}^{(-2)} \in \{(C,C),(D,D)\}\right) && (84)\\
Q_1^*\left(D,(D,C),\boldsymbol{\sigma}^{(-2)}\right) &= q_{10} \quad \left(\boldsymbol{\sigma}^{(-2)} \in \{(C,C),(D,D)\}\right) && (85)\\
Q_1^*\left(C,(D,C),\boldsymbol{\sigma}^{(-2)}\right) &= q_{11} \quad \left(\boldsymbol{\sigma}^{(-2)} \in \{(C,D),(D,C)\}\right) && (86)\\
Q_1^*\left(D,(D,C),\boldsymbol{\sigma}^{(-2)}\right) &= q_{12} \quad \left(\boldsymbol{\sigma}^{(-2)} \in \{(C,D),(D,C)\}\right) && (87)\\
Q_1^*\left(C,(D,D),\boldsymbol{\sigma}^{(-2)}\right) &= q_{13} \quad \left(\boldsymbol{\sigma}^{(-2)} \in \{(C,C),(D,D)\}\right) && (88)\\
Q_1^*\left(D,(D,D),\boldsymbol{\sigma}^{(-2)}\right) &= q_{14} \quad \left(\boldsymbol{\sigma}^{(-2)} \in \{(C,C),(D,D)\}\right) && (89)\\
Q_1^*\left(C,(D,D),\boldsymbol{\sigma}^{(-2)}\right) &= q_{15} \quad \left(\boldsymbol{\sigma}^{(-2)} \in \{(C,D),(D,C)\}\right) && (90)\\
Q_1^*\left(D,(D,D),\boldsymbol{\sigma}^{(-2)}\right) &= q_{16} \quad \left(\boldsymbol{\sigma}^{(-2)} \in \{(C,D),(D,C)\}\right) && (91)
\end{aligned}
$$

with the self-consistency condition

$$
\begin{aligned}
q_1 &> q_2\\
q_3 &< q_4\\
q_5 &> q_6\\
q_7 &< q_8\\
q_9 &> q_{10}\\
q_{11} &< q_{12}\\
q_{13} &> q_{14}\\
q_{15} &< q_{16}.
\end{aligned} \qquad (92)
$$

The solution is

$$
\begin{aligned}
q_1 = q_{13} &= \frac{1}{1-\gamma}R && (93)\\
q_2 = q_{14} &= T + \gamma R + \gamma^2 P + \frac{\gamma^3}{1-\gamma}R && (94)\\
q_3 = q_{15} &= S + \gamma R + \gamma^2 P + \frac{\gamma^3}{1-\gamma}R && (95)\\
q_4 = q_{16} &= P + \frac{\gamma}{1-\gamma}R && (96)\\
q_5 = q_9 &= R + \gamma P + \frac{\gamma^2}{1-\gamma}R && (97)\\
q_6 = q_{10} &= T + \gamma P + \gamma^2 P + \frac{\gamma^3}{1-\gamma}R && (98)\\
q_7 = q_{11} &= S + \gamma P + \gamma^2 P + \frac{\gamma^3}{1-\gamma}R && (99)\\
q_8 = q_{12} &= P + \gamma P + \frac{\gamma^2}{1-\gamma}R. && (100)
\end{aligned}
$$

For these solution, the inequalities (92) are satisfied if

$$\gamma \quad > \quad \sqrt{\frac{T-R}{R-P}}. \tag{101}$$

It should be noted that such $\gamma < 1$ exists only if $2R > T + P$. □

We remark that the condition (101) is more strict than the condition that WSLS forms a symmetric equilibrium [55]: $\gamma > \frac{T-R}{R-P}$.

5.3. All-or-None strategy

The third example of the solution of Eq. (45) is the All-or-None strategy AON_2 [51]:

$$T_1(C) = T_2(C) \quad = \quad \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}. \tag{102}$$

We can easily check that this strategy satisfies the necessary condition for symmetric equilibrium in Theorem 1. It has been known that AON_2 forms subgame perfect equilibrium [51]. A similar strategy as AON_2 was also observed in numerical simulation of evolution of cooperation [57].

Theorem 4. *When $3R - 2P - T > 0$ and $2R - 3P + S > 0$ hold, a pair of the strategy (102) forms a symmetric mutual reinforcement learning equilibrium if*

$$\gamma \quad > \quad \max\left\{ \frac{1}{2}\left(\sqrt{\frac{4T-3R-P}{R-P}} - 1 \right), \frac{1}{2}\left(\sqrt{\frac{R+3P-4S}{R-P}} - 1 \right) \right\}. \tag{103}$$

Proof. The Bellman optimality equation against the strategy (102) is

$$Q_1^*\left(C,(C,C),\boldsymbol{\sigma}^{(-2)}\right) \quad = \quad q_1 \quad \left(\boldsymbol{\sigma}^{(-2)} \in \{(C,C),(D,D)\}\right) \tag{104}$$

$$Q_1^*\left(D,(C,C),\boldsymbol{\sigma}^{(-2)}\right) \quad = \quad q_2 \quad \left(\boldsymbol{\sigma}^{(-2)} \in \{(C,C),(D,D)\}\right) \tag{105}$$

$$Q_1^*\left(C,(C,C),\boldsymbol{\sigma}^{(-2)}\right) \quad = \quad q_3 \quad \left(\boldsymbol{\sigma}^{(-2)} \in \{(C,D),(D,C)\}\right) \tag{106}$$

$$Q_1^*\left(D,(C,C),\boldsymbol{\sigma}^{(-2)}\right) \quad = \quad q_4 \quad \left(\boldsymbol{\sigma}^{(-2)} \in \{(C,D),(D,C)\}\right) \tag{107}$$

$$Q_1^*\left(C,(C,D),\boldsymbol{\sigma}^{(-2)}\right) \quad = \quad q_7 \quad \left(\boldsymbol{\sigma}^{(-2)} \in \Omega\right) \tag{108}$$

$$Q_1^*\left(D,(C,D),\boldsymbol{\sigma}^{(-2)}\right) \quad = \quad q_8 \quad \left(\boldsymbol{\sigma}^{(-2)} \in \Omega\right) \tag{109}$$

$$Q_1^*\left(C,(D,C),\boldsymbol{\sigma}^{(-2)}\right) \quad = \quad q_{11} \quad \left(\boldsymbol{\sigma}^{(-2)} \in \Omega\right) \tag{110}$$

$$Q_1^*\left(D,(D,C),\boldsymbol{\sigma}^{(-2)}\right) \quad = \quad q_{12} \quad \left(\boldsymbol{\sigma}^{(-2)} \in \Omega\right) \tag{111}$$

$$Q_1^*\left(C,(D,D),\boldsymbol{\sigma}^{(-2)}\right) = q_{13} \quad \left(\boldsymbol{\sigma}^{(-2)} \in \{(C,C),(D,D)\}\right) \tag{112}$$

$$Q_1^*\left(D,(D,D),\boldsymbol{\sigma}^{(-2)}\right) = q_{14} \quad \left(\boldsymbol{\sigma}^{(-2)} \in \{(C,C),(D,D)\}\right) \tag{113}$$

$$Q_1^*\left(C,(D,D),\boldsymbol{\sigma}^{(-2)}\right) = q_{15} \quad \left(\boldsymbol{\sigma}^{(-2)} \in \{(C,D),(D,C)\}\right) \tag{114}$$

$$Q_1^*\left(D,(D,D),\boldsymbol{\sigma}^{(-2)}\right) = q_{16} \quad \left(\boldsymbol{\sigma}^{(-2)} \in \{(C,D),(D,C)\}\right) \tag{115}$$

with the self-consistency condition

$$\begin{aligned} q_1 &> q_2 \\ q_3 &< q_4 \\ q_7 &< q_8 \\ q_{11} &< q_{12} \\ q_{13} &> q_{14} \\ q_{15} &< q_{16}. \end{aligned} \tag{116}$$

The solution is

$$q_1 = q_{13} = \frac{1}{1-\gamma}R \tag{117}$$

$$q_2 = q_{14} = T + \gamma P + \gamma^2 P + \frac{\gamma^3}{1-\gamma}R \tag{118}$$

$$q_3 = q_7 = q_{11} = q_{15} = S + \gamma P + \gamma^2 P + \frac{\gamma^3}{1-\gamma}R \tag{119}$$

$$q_4 = q_{16} = P + \frac{\gamma}{1-\gamma}R \tag{120}$$

$$q_8 = q_{12} = P + \gamma P + \frac{\gamma^2}{1-\gamma}R. \tag{121}$$

For these solution, the inequalities (116) are satisfied if

$$(R-P)\gamma^2 + (R-P)\gamma - (T-R) > 0 \tag{122}$$

and

$$(R-P)\gamma^2 + (R-P)\gamma - (P-S) > 0, \tag{123}$$

which are equivalent to Eq. (103) for $\gamma \geq 0$. It should be noted that such $\gamma < 1$ exists only if $3R - 2P - T > 0$ and $2R - 3P + S > 0$. □

6. Optimality in longer memory

In previous sections, we investigated symmetric equilibrium of mutual reinforcement learning when both players use memory-two strategies, and obtained three examples of deterministic strategies forming symmetric equilibrium. A natural question is "Do these strategies forming symmetric equilibrium

in memory-two reinforcement learning also form symmetric equilibrium of mutual reinforcement learning of longer memory strategies?". In this section, we show that the answer to this question is "yes".

We first prove the following theorem.

Theorem 5. *Let T_{-a} be a memory-n' strategy of player $-a$. Let T_a^* be the optimal strategy of player a against T_{-a} when player a use reinforcement learning of memory-n' strategies. When player a use reinforcement learning of memory-n strategies with $n > n'$ to obtain the optimal strategy $\check{T}_a^*$ against T_{-a}, then $\check{T}_a^* = T_a^*$.*

Proof. When player $-a$ use memory-n' strategy and player a use memory-n reinforcement learning with $n > n'$, the Bellman optimality equation (5) becomes

$$\begin{aligned} & Q_a^*\left(\sigma_a, \left[\boldsymbol{\sigma}^{(-m)}\right]_{m=1}^{n}\right) \\ = & \sum_{\sigma_{-a}} r_a\left(\boldsymbol{\sigma}\right) T_{-a}\left(\sigma_{-a}|\left[\boldsymbol{\sigma}^{(-m)}\right]_{m=1}^{n'}\right) \\ & +\gamma \sum_{\sigma_{-a}} T_{-a}\left(\sigma_{-a}|\left[\boldsymbol{\sigma}^{(-m)}\right]_{m=1}^{n'}\right) \max_{\hat{\sigma}} Q_a^*\left(\hat{\sigma}, \boldsymbol{\sigma}, \left[\boldsymbol{\sigma}^{(-m)}\right]_{m=1}^{n-1}\right). \end{aligned} \tag{124}$$

Then, we find that the right-hand side does not depend on $\boldsymbol{\sigma}^{(-n)}$, and therefore

$$Q_a^*\left(\sigma_a, \left[\boldsymbol{\sigma}^{(-m)}\right]_{m=1}^{n}\right) = Q_a^*\left(\sigma_a, \left[\boldsymbol{\sigma}^{(-m)}\right]_{m=1}^{n-1}\right) \tag{125}$$

Then, the Bellman optimality equation becomes

$$\begin{aligned} & Q_a^*\left(\sigma_a, \left[\boldsymbol{\sigma}^{(-m)}\right]_{m=1}^{n-1}\right) \\ = & \sum_{\sigma_{-a}} r_a\left(\boldsymbol{\sigma}\right) T_{-a}\left(\sigma_{-a}|\left[\boldsymbol{\sigma}^{(-m)}\right]_{m=1}^{n'}\right) \\ & +\gamma \sum_{\sigma_{-a}} T_{-a}\left(\sigma_{-a}|\left[\boldsymbol{\sigma}^{(-m)}\right]_{m=1}^{n'}\right) \max_{\hat{\sigma}} Q_a^*\left(\hat{\sigma}, \boldsymbol{\sigma}, \left[\boldsymbol{\sigma}^{(-m)}\right]_{m=1}^{n-2}\right). \end{aligned} \tag{126}$$

By repeating the same argument until the length of memory decreases to n', we find that

$$Q_a^*\left(\sigma_a, \left[\boldsymbol{\sigma}^{(-m)}\right]_{m=1}^{n}\right) = Q_a^*\left(\sigma_a, \left[\boldsymbol{\sigma}^{(-m)}\right]_{m=1}^{n'}\right), \tag{127}$$

which implies that $\check{T}_a^* = T_a^*$. □

That is, when the opponent $-a$ uses a memory-two strategy T_{-a}, and player a learns the optimal memory-n strategy with $n \geq 2$ against T_{-a}, then, such optimal strategy is memory-two. Similarly, when the opponent $-a$ uses a memory-one strategy, and player a learns the optimal memory-n strategy with $n \geq 1$, then, such optimal strategy is memory-one.

This theorem results in the following corollary.

Corollary 2. *A mutual reinforcement learning equilibrium obtained by memory-n' reinforcement learning is also a mutual reinforcement learning equilibrium obtained by memory-n reinforcement learning with $n > n'$.*

Therefore, the strategies (46) and (74) in the previous section also form mutual reinforcement learning equilibria even if players use memory-n reinforcement learning with $n > 2$. Similarly, the (memory-one) Grim strategy and the (memory-one) WSLS strategy still form mutual reinforcement learning equilibria even if players use memory-two reinforcement learning, since it has been known that Grim and WSLS form memory-one mutual reinforcement learning equilibria, respectively [55]. We remark that this property is similar to that of Nash equilibrium in finite automaton selection game, which claims that two automata must have an equal number of states in equilibria [8].

7. Conclusion

In this paper, we investigated symmetric equilibrium of mutual reinforcement learning when both players use memory-two deterministic strategies in the repeated prisoners' dilemma game. First, we find that the structure of the optimal strategies is constrained by the Bellman optimality equation (Proposition 1). Then, we find a necessary condition for deterministic symmetric equilibrium of mutual reinforcement learning (Theorem 1). Furthermore, we provided three examples of memory-two deterministic strategies which form symmetric mutual reinforcement learning equilibrium, some of which can be regarded as variants of the Grim strategy and the WSLS strategy (Theorem 2, Theorem 3 and Theorem 4). Finally, we proved that mutual reinforcement learning equilibria achieved by memory-two strategies are also mutual reinforcement learning equilibria when both players use reinforcement learning of memory-n strategies with $n > 2$ (Theorem 5).

We want to investigate whether other symmetric mutual reinforcement learning equilibria of deterministic memory-two strategies exist or not in future. For such purpose, novel methods are needed, because the number of strategies is quite large. Furthermore, extension of our analysis to (i) asymmetric equilibrium and (ii) mixed strategies is also a subject of future work. Ultimately, if we would develop some method to find all equilibria in all length of memory n, analysis of mutual reinforcement learning equilibria is completed.

Acknowledgement

We thank Genki Ichinose and Mashiho Mukaida for useful discussions. This study was supported by JSPS KAKENHI Grant Number JP20K19884 and Inamori Research Grants.

Appendix A. Derivation of Eq. (4)

First we introduce the notation

$$T\left(\boldsymbol{\sigma}|\left[\boldsymbol{\sigma}^{(-m)}\right]_{m=1}^{n}\right) \;:=\; \prod_{a=1}^{2} T_a\left(\sigma_a|\left[\boldsymbol{\sigma}^{(-m)}\right]_{m=1}^{n}\right). \tag{A.1}$$

We remark that the joint probability distribution of the action profiles $\boldsymbol{\sigma}^{(\mu)}$, $\cdots$, $\boldsymbol{\sigma}^{(0)}$ given $\left[\boldsymbol{\sigma}^{(-m)}\right]_{m=1}^{n}$ is described as

$$P\left(\boldsymbol{\sigma}^{(\mu)},\cdots,\boldsymbol{\sigma}^{(0)}\,\middle|\,\left[\boldsymbol{\sigma}^{(-m)}\right]_{m=1}^{n}\right) \;=\; \prod_{s=0}^{\mu} T\left(\boldsymbol{\sigma}^{(s)}\middle|\,\left[\boldsymbol{\sigma}^{(s-m)}\right]_{m=1}^{n}\right). \tag{A.2}$$

The action-value function (3) is rewritten as

$$
\begin{aligned}
& Q_a\left(\sigma_a^{(0)}, \left[\boldsymbol{\sigma}^{(-m)}\right]_{m=1}^{n}\right) \\
= & \sum_{\left[\boldsymbol{\sigma}^{(s)}\right]_{s=1}^{\infty}} \sum_{\sigma_{-a}^{(0)}} \sum_{k=0}^{\infty} \gamma^k r_a\left(\boldsymbol{\sigma}^{(k)}\right) \left\{\prod_{s=1}^{\infty} T\left(\boldsymbol{\sigma}^{(s)} \middle| \left[\boldsymbol{\sigma}^{(s-m)}\right]_{m=1}^{n}\right)\right\} \\
& \times T_{-a}\left(\sigma_{-a}^{(0)} \middle| \left[\boldsymbol{\sigma}^{(-m)}\right]_{m=1}^{n}\right) \\
= & \sum_{\left[\boldsymbol{\sigma}^{(s)}\right]_{s=1}^{\infty}} \sum_{\sigma_{-a}^{(0)}} \left[r_a\left(\boldsymbol{\sigma}^{(0)}\right) + \gamma \sum_{k=0}^{\infty} \gamma^k r_a\left(\boldsymbol{\sigma}^{(k+1)}\right)\right] \\
& \times \left\{\prod_{s=1}^{\infty} T\left(\boldsymbol{\sigma}^{(s)} \middle| \left[\boldsymbol{\sigma}^{(s-m)}\right]_{m=1}^{n}\right)\right\} T_{-a}\left(\sigma_{-a}^{(0)} \middle| \left[\boldsymbol{\sigma}^{(-m)}\right]_{m=1}^{n}\right) \\
= & \sum_{\sigma_{-a}^{(0)}} r_a\left(\boldsymbol{\sigma}^{(0)}\right) T_{-a}\left(\sigma_{-a}^{(0)} \middle| \left[\boldsymbol{\sigma}^{(-m)}\right]_{m=1}^{n}\right) \\
& + \gamma \sum_{\left[\boldsymbol{\sigma}^{(s)}\right]_{s=1}^{\infty}} \sum_{\sigma_{-a}^{(0)}} \sum_{k=0}^{\infty} \gamma^k r_a\left(\boldsymbol{\sigma}^{(k+1)}\right) \left\{\prod_{s=2}^{\infty} T\left(\boldsymbol{\sigma}^{(s)} \middle| \left[\boldsymbol{\sigma}^{(s-m)}\right]_{m=1}^{n}\right)\right\} \\
& \times T_{-a}\left(\sigma_{-a}^{(1)} \middle| \left[\boldsymbol{\sigma}^{(1-m)}\right]_{m=1}^{n}\right) T_a\left(\sigma_a^{(1)} \middle| \left[\boldsymbol{\sigma}^{(1-m)}\right]_{m=1}^{n}\right) T_{-a}\left(\sigma_{-a}^{(0)} \middle| \left[\boldsymbol{\sigma}^{(-m)}\right]_{m=1}^{n}\right) \\
= & \sum_{\sigma_{-a}^{(0)}} r_a\left(\boldsymbol{\sigma}^{(0)}\right) T_{-a}\left(\sigma_{-a}^{(0)} \middle| \left[\boldsymbol{\sigma}^{(-m)}\right]_{m=1}^{n}\right) \\
& + \gamma \sum_{\sigma_a^{(1)}} \sum_{\sigma_{-a}^{(0)}} Q_a\left(\sigma_a^{(1)}, \left[\boldsymbol{\sigma}^{(1-m)}\right]_{m=1}^{n}\right) T_a\left(\sigma_a^{(1)} \middle| \left[\boldsymbol{\sigma}^{(1-m)}\right]_{m=1}^{n}\right) T_{-a}\left(\sigma_{-a}^{(0)} \middle| \left[\boldsymbol{\sigma}^{(-m)}\right]_{m=1}^{n}\right),
\end{aligned}
\tag{A.3}
$$

which is Eq. (4).

Appendix B. Derivation of Eqs. (5) and (6)

We define Q_a^* as the optimal value of Q_a, which is obtained by choosing optimal policy T_a^*. Then, Q_a^* obeys

$$
\begin{aligned}
& Q_a^*\left(\sigma_a, \left[\boldsymbol{\sigma}^{(-m)}\right]_{m=1}^{n}\right) \\
= & \sum_{\sigma_{-a}} r_a\left(\boldsymbol{\sigma}\right) T_{-a}\left(\sigma_{-a}|\left[\boldsymbol{\sigma}^{(-m)}\right]_{m=1}^{n}\right) \\
& +\gamma \sum_{\sigma_a'} \sum_{\sigma_{-a}} T_a^*\left(\sigma_a'|\,\boldsymbol{\sigma}, \left[\boldsymbol{\sigma}^{(-m)}\right]_{m=1}^{n-1}\right) T_{-a}\left(\sigma_{-a}|\left[\boldsymbol{\sigma}^{(-m)}\right]_{m=1}^{n}\right) Q_a^*\left(\sigma_a', \boldsymbol{\sigma}, \left[\boldsymbol{\sigma}^{(-m)}\right]_{m=1}^{n-1}\right) \\
\leq & \sum_{\sigma_{-a}} r_a\left(\boldsymbol{\sigma}\right) T_{-a}\left(\sigma_{-a}|\left[\boldsymbol{\sigma}^{(-m)}\right]_{m=1}^{n}\right) \\
& +\gamma \sum_{\sigma_a'} \sum_{\sigma_{-a}} T_a^*\left(\sigma_a'|\,\boldsymbol{\sigma}, \left[\boldsymbol{\sigma}^{(-m)}\right]_{m=1}^{n-1}\right) T_{-a}\left(\sigma_{-a}|\left[\boldsymbol{\sigma}^{(-m)}\right]_{m=1}^{n}\right) \max_{\hat{\sigma}} Q_a^*\left(\hat{\sigma}, \boldsymbol{\sigma}, \left[\boldsymbol{\sigma}^{(-m)}\right]_{m=1}^{n-1}\right) \\
= & \sum_{\sigma_{-a}} r_a\left(\boldsymbol{\sigma}\right) T_{-a}\left(\sigma_{-a}|\left[\boldsymbol{\sigma}^{(-m)}\right]_{m=1}^{n}\right) \\
& +\gamma \sum_{\sigma_{-a}} T_{-a}\left(\sigma_{-a}|\left[\boldsymbol{\sigma}^{(-m)}\right]_{m=1}^{n}\right) \max_{\hat{\sigma}} Q_a^*\left(\hat{\sigma}, \boldsymbol{\sigma}, \left[\boldsymbol{\sigma}^{(-m)}\right]_{m=1}^{n-1}\right). \qquad \text{(B.1)}
\end{aligned}
$$

The equality in the third line holds when

$$
\text{supp}T_a^*\left(\cdot|\left[\boldsymbol{\sigma}^{(-m)}\right]_{m=1}^{n}\right) = \arg\max_{\sigma} Q_a^*\left(\sigma, \left[\boldsymbol{\sigma}^{(-m)}\right]_{m=1}^{n}\right), \qquad \text{(B.2)}
$$

which is Eq. (6).

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] G. J. Mailath, L. Samuelson, Repeated games and reputations: long-run relationships, Oxford University Press, 2006.

[3] D. Fudenberg, E. Maskin, The folk theorem in repeated games with discounting or with incomplete information, Econometrica: Journal of the Econometric Society 54 (3) (1986) 533–554.

[4] R. J. Aumann, Rationality and bounded rationality, Games and Economic Behavior 21 (1997) 2–14.

[5] A. Neyman, Bounded complexity justifies cooperation in the finitely repeated prisoners' dilemma, Economics Letters 19 (3) (1985) 227–229.

[6] A. Rubinstein, Finite automata play the repeated prisoner's dilemma, Journal of Economic Theory 39 (1) (1986) 83–96.

[7] E. Kalai, W. Stanford, Finite rationality and interpersonal complexity in repeated games, Econometrica: Journal of the Econometric Society 56 (2) (1988) 397–410.

[8] D. Abreu, A. Rubinstein, The structure of nash equilibrium in repeated games with finite automata, Econometrica: Journal of the Econometric Society 56 (6) (1988) 1259–1281.

[9] J. S. Banks, R. K. Sundaram, Repeated games, finite automata, and complexity, Games and Economic Behavior 2 (2) (1990) 97–117.

[10] E. Ben-Porath, Repeated games with finite automata, Journal of Economic Theory 59 (1) (1993) 17–32.

[11] A. Neyman, Finitely repeated games with finite automata, Mathematics of Operations Research 23 (3) (1998) 513–552.

[12] E. Lehrer, Repeated games with stationary bounded recall strategies, Journal of Economic Theory 46 (1) (1988) 130–144.

[13] H. Sabourian, Repeated games with m-period bounded memory (pure strategies), Journal of Mathematical Economics 30 (1) (1998) 1–35.

[14] M. Barlo, G. Carmona, H. Sabourian, Repeated games with one-memory, Journal of Economic Theory 144 (1) (2009) 312–336.

[15] M. Barlo, G. Carmona, H. Sabourian, Bounded memory folk theorem, Journal of Economic Theory 163 (2016) 728–774.

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

[17] R. Boyd, J. P. Lorberbaum, No pure strategy is evolutionarily stable in the repeated prisoner's dilemma game, Nature 327 (6117) (1987) 58–59.

[18] D. Fudenberg, E. Maskin, Evolution and cooperation in noisy repeated games, The American Economic Review 80 (2) (1990) 274–279.

[19] K. G. Binmore, L. Samuelson, Evolutionary stability in repeated games played by finite automata, Journal of Economic Theory 57 (2) (1992) 278–305.

[20] M. A. Nowak, K. Sigmund, E. El-Sedy, Automata, repeated games and noise, Journal of Mathematical Biology 33 (7) (1995) 703–722.

[21] M. A. Nowak, K. Sigmund, Tit for tat in heterogeneous populations, Nature 355 (6357) (1992) 250–253.

[22] 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.

[23] C. T. Bergstrom, M. Lachmann, The red king effect: when the slowest runner wins the coevolutionary race, Proceedings of the National Academy of Sciences 100 (2) (2003) 593–598.

[24] L. A. Imhof, D. Fudenberg, M. A. Nowak, Evolutionary cycles of cooperation and defection, Proceedings of the National Academy of Sciences 102 (31) (2005) 10797–10800.

[25] A. Szolnoki, M. Perc, G. Szabó, Phase diagrams for three-strategy evolutionary prisoner's dilemma games on regular graphs, Physical Review E 80 (5) (2009) 056104.

[26] M. Perc, J. Gómez-Gardenes, A. Szolnoki, L. M. Floría, Y. Moreno, Evolutionary dynamics of group interactions on structured populations: a review, Journal of the Royal Society Interface 10 (80) (2013) 20120997.

[27] A. J. Stewart, J. B. Plotkin, From extortion to generosity, evolution in the iterated prisoner's dilemma, Proceedings of the National Academy of Sciences 110 (38) (2013) 15348–15353.

[28] E. Kalai, E. Lehrer, Rational learning leads to nash equilibrium, Econometrica: Journal of the Econometric Society (1993) 1019–1045.

[29] D. Fudenberg, D. K. Levine, Steady state learning and nash equilibrium, Econometrica: Journal of the Econometric Society (1993) 547–573.

[30] S. Hart, A. Mas-Colell, A simple adaptive procedure leading to correlated equilibrium, Econometrica 68 (5) (2000) 1127–1150.

[31] T. Roughgarden, Twenty lectures on algorithmic game theory, Cambridge University Press, 2016.

[32] D. Kraines, V. Kraines, Pavlov and the prisoner's dilemma, Theory and Decision 26 (1) (1989) 47–79.

[33] 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.

[34] 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.

[35] M. W. Macy, A. Flache, Learning dynamics in social dilemmas, Proceedings of the National Academy of Sciences 99 (suppl 3) (2002) 7229–7236.

[36] N. Masuda, M. Nakamura, Numerical analysis of a reinforcement learning model with the dynamic aspiration level in the iterated prisoner's dilemma, Journal of Theoretical Biology 278 (1) (2011) 55–62.

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

[38] D. Fudenberg, D. K. Levine, The theory of learning in games, Vol. 2, MIT Press, 1998.

[39] I. Erev, A. E. Roth, Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria, American Economic Review 88 (4) (1998) 848–881.

[40] P. Dal Bó, Cooperation under the shadow of the future: experimental evidence from infinitely repeated games, American Economic Review 95 (5) (2005) 1591–1604.

[41] P. Dal Bó, G. R. Fréchette, The evolution of cooperation in infinitely repeated games: Experimental evidence, American Economic Review 101 (1) (2011) 411–29.

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

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

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

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

[46] 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.

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

[48] 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.

[49] J. Li, G. Kendall, The effect of memory size on the evolutionary stability of strategies in iterated prisoner's dilemma, IEEE Transactions on Evolutionary Computation 18 (6) (2013) 819–826.

[50] S. D. Yi, S. K. Baek, J.-K. Choi, Combination with anti-tit-for-tat remedies problems of tit-for-tat, Journal of Theoretical Biology 412 (2017) 1–7.

[51] C. Hilbe, L. A. Martinez-Vaquero, K. Chatterjee, M. A. Nowak, Memory-n strategies of direct reciprocity, Proceedings of the National Academy of Sciences 114 (18) (2017) 4715–4720.

[52] Y. Murase, S. K. Baek, Seven rules to avoid the tragedy of the commons, Journal of Theoretical Biology 449 (2018) 94–102.

[53] Y. Murase, S. K. Baek, Five rules for friendly rivalry in direct reciprocity, Scientific Reports 10 (2020) 16904.

[54] M. Ueda, Memory-two zero-determinant strategies in repeated games, Royal Society Open Science 8 (5) (2021) 202186.

[55] Y. Usui, M. Ueda, Symmetric equilibrium of multi-agent reinforcement learning in repeated prisoner's dilemma, Applied Mathematics and Computation 409 (2021) 126370.

[56] J. W. Friedman, A non-cooperative equilibrium for supergames, The Review of Economic Studies 38 (1) (1971) 1–12.

[57] C. Hauert, H. G. Schuster, Effects of increasing the number of players and memory size in the iterated prisoner's dilemma: a numerical approach, Proceedings of the Royal Society of London. Series B: Biological Sciences 264 (1381) (1997) 513–519.