Repeated Triangular Trade: Sustaining Circular Cooperation with Observation Errors

Kota Shigedomi[1], Fuuki Shigenaka[1], Tadashi Sekiguchi[2], Atsushi Iwasaki[3], and Makoto Yokoo[1]

[1] Kyushu University, Motooka 744, Fukuoka, Japan.
{shigedomi@agent., shigenaka@agent., yokoo@}inf.kyushu-u.ac.jp
[2] Kyoto University, Yoshida-Honmachi, Sakyo-ku, Kyoto, Japan.
sekiguchi@kier.kyoto-u.ac.jp
[3] University of Electro-Communications, Chofugaoka 1-5-1, Chofu, Tokyo, Japan.
iwasaki@is.uec.ac.jp

Abstract. We introduce a new fundamental problem called *triangular trade*, which is a natural extension of the well-studied prisoner's dilemma for three (or more) players. This problem deals with a situation in which players would be better off if they maintain circular cooperation, but each player has an incentive to defect and a player cannot directly punish a (seemingly) defected player. We analyze whether players can sustain such circular cooperation when they repeatedly play this game and each player observes the action of another player with some observation errors (imperfect private monitoring). We confirm that no simple strategy can constitute an equilibrium within any reasonable parameter settings when there are only two actions: "Cooperate" and "Defect." Thus, we introduce two additional actions: "Whistle" and "Punish," which can be considered as a slight modification of "Cooperate." Then, players can achieve sustainable cooperation using a simple strategy called remote punishment strategy, which constitutes an equilibrium for a wide range of parameter settings. The sufficient/necessary condition, in which it constitutes an equilibrium, is represented by a simple closed-form condition. Furthermore, this strategy can be extended to the case with four or more players.

Keywords: Repeated games, private monitoring, belief-free equilibrium

1 Introduction

The prisoner's dilemma (PD) concisely represents a ubiquitous situation where the cooperation of two players is efficient, but each player has an incentive to defect. A repeated game, where players repeatedly play the same stage game (e.g., PD) over an infinite time horizon, is a formal model that can explain why cooperation arises in long-term relationships. In this paper, we introduce a new problem called *triangular trade*, which is similar to the PD, but a player cannot directly punish a player who has (seemingly) defected. In the real world, there exist many situations where the power/influence of players is not symmetric and can be considered as one-way (e.g., a teacher to a student/parent, a TV production company to a viewer). It is also common that such one-way relations create a cycle, e.g., a teacher affects a student/parent, the student/parent affects the school attended by the student, and the school affects the teacher, or a production company affects a viewer, the viewer affects a sponsor, and the sponsor affects the production company. In such situations, as in the PD, it is quite possible that the cooperation of all the players is efficient (e.g., the teacher gives a good lecture, the student/parent donate enough money, and the school supports the teacher), but each player has an incentive to defect. The triangular trade is a problem that concisely represents such a ubiquitous situation, which can be considered as a natural extension of the PD for three or more players. Such a situation can occur in international trade among three countries, where the trade between two countries is strongly imbalanced [16]. Since a player cannot directly punish a player who has seemingly defected, obtaining sustainable cooperation seems difficult (or even impossible) when a player can only imperfectly observe the actions of the other players. To the best of our knowledge, we are the first to analyze this simple but fundamental problem in repeated games with imperfect private monitoring.

A repeated game has received considerable attention in AI, multi-agent systems, and economics literature. The case of perfect monitoring, where each player can observe other players' actions, is now well understood. There is also a large body of literature on the *imperfect monitoring* case, where players' actions are only imperfectly observed through some signals. Such imperfect monitoring cases are further classified into *public* and *private monitoring* cases. If *all* players observe the same set of signals that imperfectly indicate players' actions, we have an *imperfect public monitoring* case. An example is the PD game with action-errors, investigated by Nowak and Sigmund [18]. In contrast, suppose that each player observes her opponent's action with some observation errors. Assume that each player chooses "Cooperate" (C) or "Defect" (D), and a signal, which determines a player's outcome, can be either good (g) or bad (b). If the opponent plays C, a player usually observes g, but she may observe b with a small probability. An important feature of this model is that a player's observation is her private information that is not known to the opponent. This is an example of *imperfect private monitoring*, where each player privately receives signals about the actions of other players. In private monitoring, verifying an equilibrium becomes hard since we need to check that no player has an incentive to deviate under any possible belief she might have on the past histories of other players. To overcome this difficulty, a special type of equilibrium called *belief-free* equilibrium is identified, where checking whether a profile of strategies forms such an equilibrium becomes more tractable [9, 21]. Also, what kinds of cooperative relations can be sustainable in the repeated PD is examined [8].

In this paper, we first show that when there are only two actions, i.e., C and D (and their associated observations), there exists no simple strategy that constitutes a belief-free equilibrium. We confirm this fact by exhaustively enumerating all simple strategies. Thus, we consider adding new actions, which we call “Whistle” and “Punish.” These actions are similar to C; they are dominated by D. Thus, adding them is irrelevant in a one-shot game, i.e., they do not affect equilibria. Introducing such an action is not interesting with perfect or imperfect public monitoring, since it is well-known that cooperative relations are sustainable without introducing such an action due to the celebrated folk theorem [12, 11]. With imperfect private monitoring, introducing an action that can severely punish other players can be effective even if the action is dominated by another action, i.e., the equilibria of a repeated game may significantly change if the added action changes the players’ minimax values. We emphasize that our argument is *not* based on this logic because these actions are rather mildly spiteful actions that do not change the minimax values.

To our surprise, it turns out that by adding these actions, players can achieve sustainable cooperation using a very simple strategy called Remote Punishment (RP) strategy, which constitutes a belief-free equilibrium in a wide range of parameter settings. We obtain a simple closed-form sufficient/necessary condition, in which it constitutes a belief-free equilibrium. Furthermore, we show how to extend this strategy to the case with four or more players.

2 Model

2.1 Repeated triangular trade with imperfect private monitoring

Let us describe the basic model of the triangular trade with three players. There exists three players $N = \{0, 1, 2\}$. Each player $i \in N$ repeatedly plays the same stage game over an infinite horizon $t = 0, 1, 2, \ldots$. In each period, player i takes some action a_i from a finite set A. Assume an action profile in that period is $\boldsymbol{a} = (a_0, a_1, a_2) \in A^3$. Then, her expected payoff in that period is given by stage game payoff function $u_i(\boldsymbol{a})$. In the triangular trade, we assume that the stage game payoff of player i depends only on her own action a_i and the action of player $i-1$, i.e., a_{i-1}. Throughout this paper, when we write player $i \pm k$ ($k \in \mathbb{N}$), it means player $i \pm k \bmod 3$.[4]

In the basic model, we assume that $A = \{C, D\}$ and u_i is given in Table 1, where $0 < c < 1$. In the triangular trade, player i makes a product and delivers it to player $i + 1$. Action C means that a player exerts adequate effort in the production (which incurs cost c), while action D means that a player exerts no effort at all (which incurs no cost). By receiving a product made with adequate effort, a player obtains benefit 1, while by receiving a product made with no effort, a player obtains no benefit. Note that this game has a similar characteristic to the PD. Here, C is dominated by D. Thus, in the one-shot game, the dominant strategy equilibrium is that all players play D and their utilities are 0. However, if they play C, their utilities are $1-c > 0$. The triangular trade can be considered as a natural extension of the PD for three (or more) players. More specifically, a typical domain where a PD like situation would occur is *mutual aid*; each player has her own task, which can be done better with the help of another player, but

[4] The same applies to action $a_{i\pm k}$ or state $\theta_{i\pm k}$.

Table 1. Stage game payoff (two actions)

$a_i \backslash a_{i-1}$	C	D
C	$1-c$	$-c$
D	1	0

a player obtains no merit by helping another. Assume a similar situation with three players, where each task requires at most two players, then a natural and efficient way is to maintain circular cooperation. Here, we have exactly the same problem setting as the triangular trade. Also, the situation can be generalized to the case with four or more players (Section 5).

Within each period, player i observes her private signal $\omega_i \in \Omega$ that is related to player $i-1$'s action. In the triangular trade, $\Omega = \{g, b\}$. Observation g means that the delivered product from $i-1$ has high quality and b means that it has low quality. Let $\boldsymbol{\omega} = (\omega_0, \omega_1, \omega_2) \in \Omega^3$ denote the profile of the private signals for all players. Let $o(\omega_i \mid a_{i-1})$ denote the marginal distribution of ω_i given player $i-1$'s action a_{i-1}. The signals are independent, i.e., the probability that players receive the profile of private signals $\boldsymbol{\omega}$ when players take $\boldsymbol{a}$ is given as $o(\boldsymbol{\omega} \mid \boldsymbol{a}) = \prod_{i \in N} o(\omega_i \mid a_{i-1})$.

We assume *nearly-perfect* monitoring. When a player chooses C (or D), we assume that the "correct" signal is g (or b). We assume a player receives a correct signal with high probability q but she receives a wrong signal with small probability $1-q$. Also, we assume no player can infer which action was taken (or not taken) by another player for sure; each signal $\omega_i \in \Omega$ occurs with a positive probability for any $a_{i-1} \in A$ (*full-support assumption*).

Player i's *realized* payoff, which is determined by her own action and signal, is denoted as $\pi_i(a_i, \omega_i)$. Hence, her expected payoff is given by $\sum_{\omega_i \in \Omega} \pi_i(a_i, \omega_i) \cdot o(\omega_i \mid a_{i-1})$. The product can occasionally have low (or high) quality even if the player exerts adequate effort (or no effort). It is natural to assume that the benefit of the product is solely determined by its quality. We assume this expected value of the realized payoff is identical to stage game payoff $u_i(\boldsymbol{a})$. The particular values of the realized payoffs are not important for analyzing equilibria since their expected value, which is equal to $u_i(\boldsymbol{a})$, depends only on the action profile $\boldsymbol{a}$. Thus, in this paper, we do not specify the particular values of the realized payoffs. This model is standard in the literature of repeated games with private monitoring [17].

The stage game is repeatedly played over an infinite horizon. Player i's expected discounted payoff from a sequence of action profiles $\boldsymbol{a}^0, \boldsymbol{a}^1, \ldots$ is $\sum_{t=0}^{\infty} \delta^t u_i(\boldsymbol{a}^t)$, with discount factor $\delta \in (0,1)$. The (expected) discounted *average payoff* (payoff per period) is defined as $(1-\delta) \sum_{t=0}^{\infty} \delta^t u_i(\boldsymbol{a}^t)$.

2.2 Strategy representation and equilibrium concept

For player i, the set of her private histories at period t is $H_i^t := (A \times \Omega)^t$. Each element $h_i^t = (a_i^0, \omega_i^0, ..., a_i^{t-1}, \omega_i^{t-1}) \in H_i^t$ represents the sequence of her actions and observation profiles until the end of period $t-1$. H_i^0 is interpreted as a singleton, which represents a (dummy) initial history. A (pure) strategy for player i is represented as function $s_i : H_i \to A$, which returns the action that player i should choose at period t given her history h_i^t. Here, H_i is all the possible histories of i, i.e., $\bigcup_{t\geq 0} H_i^t$. Let $\boldsymbol{s} = (s_i, \boldsymbol{s}_{-i})$ denote the profile of strategies, where s_i is i's strategy and $\boldsymbol{s}_{-i}$ is the profile of the strategies of the other players. Let $E_i(\boldsymbol{s})$ denote player i's discounted average payoff when all the players act based on strategy profile $\boldsymbol{s}$. We say s_i is a best response to $\boldsymbol{s}_{-i}$ if for any possible strategy s_i' of player i, $E_i((s_i, \boldsymbol{s}_{-i})) \geq E_i((s_i', \boldsymbol{s}_{-i}))$ holds.

A standard equilibrium concept in repeated games is a *sequential equilibrium*, which is a refinement of a subgame perfect equilibrium as well as a perfect Bayesian equilibrium [14]. In a private monitoring setting, profile of strategies $\boldsymbol{s}$ is a sequential equilibrium if for each $i \in N$, for any t, for any history $h_i^t \in H_i^t$, and a possible belief reached after observing h_i^t, acting according to s_i (for given history h_i^t) is a best response under the belief.

A Finite-State Automaton (FSA) is a popular approach for concisely representing a strategy in an infinitely repeated game. Player i's FSA M_i is defined by $\langle \Theta_i, \hat{\theta}_i, f_i, T_i \rangle$, where Θ_i is a set of states, $\hat{\theta}_i \in \Theta_i$ is an initial state, $f_i : \Theta_i \to A$ determines the action choice in each state, and $T_i : \Theta_i \times \Omega \to \Theta_i$ specifies a deterministic state transition. Specifically, $T_i(\theta_i^t, \omega_i^t)$ returns next state θ_i^{t+1} when the current state is θ_i^t and player i's private signal is ω_i^t. For M_i and h_i^t, the action to choose in period t is defined as $f_i(\theta_i^t)$, where θ_i^t is the state reached after history h_i^t.

An FSA without specification of the initial state, i.e., $m_i = \langle \Theta_i, f_i, T_i \rangle$, is a *Finite-State preAutomaton* (pre-FSA). $(m_i, \hat{\theta}_i)$ denotes an FSA obtained by m_i, where the initial state is $\hat{\theta}_i$. Let $\boldsymbol{M} = (M_i)_{i \in N}$ denote a profile of FSAs. Figure 1 shows an example of a pre-FSA. Each node represents a state and a direct link represents a state transition according to an observation.

For M_i, let $\Theta_i^t \subseteq \Theta_i$ denote a set of states reachable in period t. By the full-support assumption, Θ_i^t is determined independently from the strategies of other players.

Now, we are ready to define a belief-free equilibrium.

Definition 1 (Belief-free equilibrium). *We say $\boldsymbol{M}$ is a belief-free equilibrium if for all t, for all $\boldsymbol{\theta} = (\theta_i)_{i \in N} \in \prod_{i \in N} \Theta_i^t$, and for all $i \in N$, (m_i, θ_i) is a best response when player $j \neq i$ is going to behave based on (m_j, θ_j) .*

Note that we are not restricting the possible strategy spaces of players (i.e., we are *not* assuming that players can only use FSAs). The requirement that (m_i, θ_i) is a best response implies that her discounted average payoff cannot be improved even if she uses a very sophisticated strategy, which cannot be represented by an FSA.

It is obvious that a belief-free equilibrium is a special case of a sequential equilibrium, since a sequential equilibrium requires that the strategy of each player be a best response under *all beliefs that are reachable*, while a belief-free equilibrium requires that her strategy be a best response under *all beliefs including the unreachable ones.*

Table 2. Stage game payoff (four actions)

$a_i \backslash a_{i-1}$	C	D	W	P
C	$1-c$	$-c$	$1-y-c$	$1-z-c$
D	1	0	$1-y$	$1-z$
W	$1-c$	$-c$	$1-y-c$	$1-z-c$
P	$1-c$	$-c$	$1-y-c$	$1-z-c$

3 Game with additional actions and observations

We exhaustively generated all small FSAs (each of which has at most three states) and confirmed that none of them constitutes a belief-free equilibrium under any reasonable parameter settings, except for a trivial strategy that simply plays D forever. More specifically, we checked parameter settings in which $0.1 \leq \delta < 1$ (in increments of 0.2), $0.55 \leq q < 1$ (in increments of 0.1), and $0.1 \leq c < 1$ (in increments of 0.1).

Thus, we consider a slightly modified game with two additional actions and observations. Here, we assume a player can slightly modify action C. More specifically, the player actually exerts adequate effort to produce a product, but she intentionally damages the product such that the benefit for the receiver is reduced and the receiver notices (with high probability) that the producer intentionally did so. There are two additional actions W ("Whistle") and P ("Punish"), and two associated observations w and p. The stage game payoff is given in Table 2. Here, doing W or P incurs cost c (as doing C). When player $i-1$ plays W (or P), player i's benefit is reduced by y (or z) compared to the case where player $i-1$ plays C. We assume $0 \leq y, z \leq 1$, i.e., P (or W) is a relatively mild spiteful action; it is weaker than (or at most equal to) making the product totally useless.

For actions C, D, W, and P, their "correct" signals are g, b, w, and p, respectively. Then, the observation probability $o(\omega_i \mid a_{i-1})$ is q when ω_i is the correct signal, and $e = (1-q)/3$ when ω_i is an incorrect signal. We assume $1/4 < q < 1$, i.e., the correct signal is most likely.

4 Remote Punishment (RP) strategy

We identify a very simple strategy that can constitute a belief-free equilibrium with a wide range of parameter settings. We call it Remote Punishment (RP) strategy. It is a kind of "reactive" strategy, in which the action in the current period is determined by the signal in the previous period. The signal-action mapping is given in Table 3. In words, as long as player i observes g, she plays C. When she observes b, i.e., player $i-1$ seems to play D, then she informs this fact to player $i+1$ by playing W. When she observes w, i.e., player $i-1$ seems to ask her to punish $i+1$, she plays P. Finally, when she observes p, i.e., player $i-1$ seems to punish her, she tolerates the punishment and plays C.

The pre-FSA for the RP strategy is given in Figure 1. There are three states S_C, S_W, S_P. In each state S_{a_i} (where $a_i \in \{C, W, P\}$), player i plays the speci-

Table 3. Signal-action mapping of RP

Previous signal	Current action
g	C
b	W
w	P
p	C

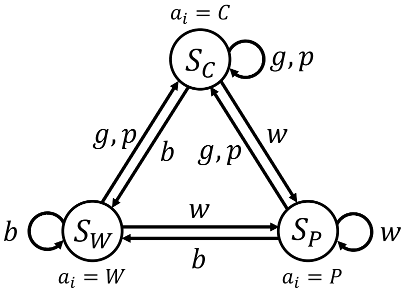

Fig. 1. Pre-FSA of RP

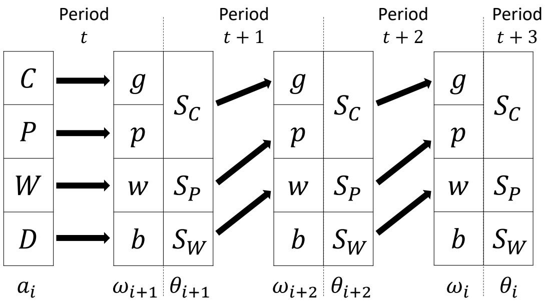

Fig. 2. Effect of action selection for probability distribution of states

fied action a_i. The initial state is S_C. We illustrate in Figure 2 how the action selection in period t of player i affects the possible states of players $i+1$, $i+2$, and i for periods $t+1$, $t+2$, and $t+3$, respectively. The thick arrow connects an action/state and its "correct" signal, which is observed with probability q. For example, when the action of player i in period t is C, the observation of player $i+1$ is g with probability q (and the probabilities of the other "wrong" signals are e). Thus, the state of player $i+1$ in period $t+1$ is S_c with probability $q+e = 1-2e$, and the probabilities of the other states are e.

The following theorem characterizes the condition where the RP strategy constitutes a belief-free equilibrium.

Theorem 1. *The sufficient and necessary condition in which the RP strategy constitutes a belief-free equilibrium is:*

$$\delta^2(1-4e)^2 z \geq c. \tag{1}$$

This condition is intuitively natural; it says that sustainable cooperation is more likely to be established when (i) the cost of cooperation c is small, (ii) the players are patient (i.e., δ is large), (iii) the error probability (e) is small, and (iv) the punishment z is large. Assume c is 0.2 and δ is 0.9. Also, assume the punishment z must be at most 1, i.e., any severe punishment, which is stronger than making the product totally useless, is not possible. Then, $(1-4e)$ must be approximately more than 1/2, which implies that e is less than 1/8 (and q is more than 5/8). Note that the value of y, i.e., the amount of reduced benefit for player i when player $i-1$ plays W, is irrelevant to this condition. Thus, it can be 0.

To prove this theorem, we need to show that player i has no incentive to deviate from the RP strategy regardless of the current states of the other players. Although there exist infinitely many possible deviations, by Proposition 12.2.3 in [17], it is sufficient to check a finite number of deviations, each of which chooses a different action only once, and immediately returns to the original strategy. This property is called one-shot deviation principle or one-deviation property [17].

Since all players use the same strategy, it is sufficient to check that there exists no profitable deviation for player 0. We utilize the following lemmas.

Lemma 1. *Let $\boldsymbol{\gamma}_{i,t}$ (and $\boldsymbol{\gamma}'_{i,t}$) denote the probability distribution of the state of player i at period t, i.e., it is a three element vector $\begin{pmatrix}\gamma_{S_C} & \gamma_{S_P} & \gamma_{S_W}\end{pmatrix}$. Also, for given $\boldsymbol{\gamma}_{i,t}$ (or $\boldsymbol{\gamma}'_{i,t}$), let $\boldsymbol{\gamma}_{i+k,t+k}$ (or $\boldsymbol{\gamma}'_{i+k,t+k}$) denote the probability distribution of the state of player $i+k$ at period $t+k$ when all players follow the RP strategy. For all $k \geq 2$, for all $\boldsymbol{\gamma}_{i,t}, \boldsymbol{\gamma}'_{i,t}$, the following condition holds.*

$$\boldsymbol{\gamma}_{i+k,t+k} = \boldsymbol{\gamma}'_{i+k,t+k}.$$

In words, for $k \geq 2$, the probability distribution of player $i+k$'s state in period $t+k$ is identical regardless of the probability distribution of player i's state in period t.

Proof (Proof of Lemma 1). Let X denote the following matrix:

$$X = \begin{pmatrix} 1-2e & e & e \\ 1-2e & e & e \\ 2e & 1-3e & e \end{pmatrix},$$

i.e., it represents transition probabilities from one state to another. Then, $\boldsymbol{\gamma}_{i+1,t+1}$ is given by $\boldsymbol{\gamma}_{i,t}X$, and $\boldsymbol{\gamma}_{i+k,t+k}$ is given by $\boldsymbol{\gamma}_{i,t}X^k$. X^2 is calculated as follows:

$$X^2 = \begin{pmatrix} 4e^2-3e+1 & -4e^2+2e & e \\ 4e^2-3e+1 & -4e^2+2e & e \\ 4e^2-3e+1 & -4e^2+2e & e \end{pmatrix}.$$

Since all the row vectors in X^2 are identical, and for each row vector of X, the sum of its elements is 1, for any $k \geq 2$, all the row vectors in X^k are identical. Actually, for any $k \geq 2$, $X^k = X^2$ holds. Thus, for any $k \geq 2$, for all $\boldsymbol{\gamma}_{i,t}$ and $\boldsymbol{\gamma}'_{i,t}$, $\boldsymbol{\gamma}_{i+k,t+k} = \boldsymbol{\gamma}'_{i+k,t+k}$ holds. □

Let $V^{\boldsymbol{\theta}}$ denote the average discounted payoff of player 0 when all players follow the RP strategy and start from $\boldsymbol{\theta}$.

Lemma 2. *For all $\theta_1, \theta_2 \in \{S_C, S_W, S_P\}$, the following condition holds:*

$$V^{(S_C,\theta_1,\theta_2)} = V^{(S_W,\theta_1,\theta_2)} = V^{(S_P,\theta_1,\theta_2)}.$$

In words, the discounted average payoff of player 0 is identical regardless of her current state. This is a necessary condition for a belief-free equilibrium.

Proof (Proof Sketch of Lemma 2). Assume the current period is t. Since all actions C, W, P have the same cost, the expected reward of period t must be identical. The expected reward in period $t+1$ is determined independently from the action of period t. From Lemma 1, the probability distribution of the states of player 2 in period $t+2$ and thereafter must be identical regardless of the state of player 0 at period t. Thus, the expected reward at period $t+2$ and thereafter must be identical. □

Lemma 3. *Assume player i plays action $a_i \in A$ (which might be different from the one specified by the RP strategy) at period t and acts according to the RP strategy thereafter. Then, for all $k \geq 3$, the conditional probability distributions of the states of player $i+k$ at period $t+k$ are identical regardless of a_i.*

Proof (Proof of Lemma 3). Let $\boldsymbol{\gamma}_C, \boldsymbol{\gamma}_D, \boldsymbol{\gamma}_W$, and $\boldsymbol{\gamma}_P$ denote the probability distributions of the states of player $i+1$ at period $t+1$, assuming player i plays C, D, W, and P at period t, respectively. Then, the probability distribution of player $i+k$ at period $t+k$, assuming player i plays C (or D, W, P) at period t, is given as $\boldsymbol{\gamma}_C X^{k-1}$ (or $\boldsymbol{\gamma}_D X^{k-1}$, $\boldsymbol{\gamma}_W X^{k-1}$, $\boldsymbol{\gamma}_P X^{k-1}$). From Lemma 1, these distributions are identical when $k-1 \geq 2$. □

Now, we are ready to prove Theorem 1.

Proof (Proof of Theorem 1). Assume player 0 deviates in period t (and returns to the RP strategy after $t+1$). It is sufficient to compare the following two values: (i) the cost of player 0 at period t (which is determined by her chosen action a_0 in period t), and (ii) the benefit of player 0 at period $t+2$ (which is determined by a_2 in period $t+2$). This is because, at period t, all the players except 0 follow the strategy. Thus, the benefit of player 0 is unchanged (only her cost can vary). At period $t+1$, the action of player 2 is unchanged. At period $t+2$, the benefit of player 0 is affected by the action of player 2. At and after period $t+3$, the probabilistic distributions of the states of the other players are the same (Lemma 3).

Let us compare these values for possible deviations. When $(\theta_1, \theta_2) = (S_C, S_C)$, for the deviation from C to D at period t, the payoff increases by c since she exerts no effort. Let us examine the decreased amount of player 0's payoff at period $t+2$. The probability distribution of player 2's states after two periods can be represented as follows: (i) $\begin{pmatrix}1-2e & e & e\end{pmatrix} X$ when player 0 chooses C, (ii) $\begin{pmatrix}2e & e & 1-3e\end{pmatrix} X$ when player 0 deviates to D. Then, the difference is given as:

$$\begin{aligned}
&\delta^2 \begin{pmatrix}2e & e & 1-3e\end{pmatrix} X \begin{pmatrix}1 & 1-z & 1-y\end{pmatrix}^{\mathrm{T}} \\
&\quad - \delta^2 \begin{pmatrix}1-2e & e & e\end{pmatrix} X \begin{pmatrix}1 & 1-z & 1-y\end{pmatrix}^{\mathrm{T}} \\
&= \delta^2 \begin{pmatrix}-(1-4e) & 0 & 1-4e\end{pmatrix} X \begin{pmatrix}1 & 1-z & 1-y\end{pmatrix}^{\mathrm{T}} \\
&= -\delta^2 (1-4e)^2 z.
\end{aligned}$$

Thus, the incentive constraint is given as Inequality (1). For the deviation from W or P to D, the results are the same as above. Also, from Lemma 2, there is no incentive for any deviation among C, P, and W. The proof for the case when $(\theta_1, \theta_2) \neq (S_C, S_C)$ is similar to the above. □

The following theorem derives the average discounted payoff for the RP strategy.

Theorem 2. *When all players play the RP strategy, the average discounted payoff, i.e.,* $V^{(S_C, S_C, S_C)}$*, is given as follows:*

$$(1-c) - \delta^2 e(1-4e)z - \delta e y - \delta e z. \tag{2}$$

Here, $(1-c)$ is the ideal payoff when players keep on cooperating and no errors occur. Due to the errors, the average discounted payoff is reduced to some extent.

Proof (Proof of Theorem 2). As we showed in the proof of Lemma 1, for all $k \geq 2$, $X^k = X^2$ holds. The probability distribution of the states of player 2 at period t is given as follows (we assume X^0 is an identity matrix):

$$\begin{pmatrix}1 & 0 & 0\end{pmatrix} X^t.$$

As long as player 0 follows the RP strategy, the cost of her action is c. Also, the baseline benefit of player 0, which is determined by the action of player 2, is 1. Let us examine how the decreased amount of player 0's benefit affects her average discounted benefit. This amount is given as follows.

$$\begin{aligned}
&(1-\delta)\sum_{t=0}^{\infty}\delta^t \begin{pmatrix}1 & 0 & 0\end{pmatrix} X^t \begin{pmatrix}0 & z & y\end{pmatrix}^{\mathrm{T}} \\
&= (1-\delta)\delta \begin{pmatrix}1 & 0 & 0\end{pmatrix} X \begin{pmatrix}0 & z & y\end{pmatrix}^{\mathrm{T}} \\
&\quad + (1-\delta)\sum_{t=2}^{\infty}\delta^t \begin{pmatrix}1 & 0 & 0\end{pmatrix} X^2 \begin{pmatrix}0 & z & y\end{pmatrix}^{\mathrm{T}} \\
&= (1-\delta)\delta \begin{pmatrix}1-2e & e & e\end{pmatrix} \begin{pmatrix}0 & z & y\end{pmatrix}^{\mathrm{T}} \\
&\quad + \delta^2 \begin{pmatrix}4e^2 - 3e + 1 & -4e^2 + 2e & e\end{pmatrix} \begin{pmatrix}0 & z & y\end{pmatrix}^{\mathrm{T}} \\
&= \delta e(y+z) + \delta^2 e(1-4e)z
\end{aligned}$$

Thus, the average discounted payoff is given as Equation (2). □

5 Extension to the case with four or more players

Let us extend the RP strategy for the case of $n \geq 4$. There exists a set of players: $N = \{0, 1, 2, \ldots, n-1\}$ $(n \geq 4)$. In this section, when we write player $i \pm k$, it means player $i \pm k \bmod n$. The same applies to action $a_{i \pm k}$ or state $\theta_{i \pm k}$. Possible actions are $A = \{C, D, W_1, \ldots, W_{n-2}, P\}$. The cost of action D is 0. For other actions, its cost is c. Signals are $\Omega = \{g, b, w_1, \ldots, w_{n-2}, p\}$, each of

Table 4. Stage game payoff of player i ($n \geq 4$)

$a_i \backslash a_{i-1}$	C	D	W_j	P
C	$1-c$	$-c$	$1-y_j-c$	$1-z-c$
D	1	0	$1-y_j$	$1-z$
W_k	$1-c$	$-c$	$1-y_j-c$	$1-z-c$
P	$1-c$	$-c$	$1-y_j-c$	$1-z-c$

Table 5. Signal-action mapping of RP ($n \geq 4$)

Previous signal	Current action
g	C
b	W_1
w_1	W_2
$\dots$	$\dots$
w_{n-3}	W_{n-2}
w_{n-2}	P
p	C

which corresponds to the correct signal of $C, D, W_1, \dots, W_{n-2}, P$, respectively. The associated benefits are $1, 0, 1-y_1, \dots, 1-y_{n-2}, 1-z$, respectively. Player i observes signal ω_i, which is determined by player $i-1$'s action. We assume signals are independent and private. The stage game payoff is defined in Table 4. Player i's stage game payoff depends only on a_i and a_{i-1}. We assume $0 \leq y_1, \dots, y_{n-2}, z \leq 1$. The observation probability $o(\omega_i \mid a_{i-1})$ is given as q when ω_i is the correct signal, or $e = (1-q)/n$ when ω_i is an incorrect signal. We assume $1/(n+1) < q < 1$, i.e., the correct signal is most likely.

The RP strategy for $n \geq 4$ is also a "reactive" strategy, in which the action in the current period is determined by the signal in the previous period. The signal-action mapping is given in Table 5. When player $i+1$ observes b, she notifies this fact by playing W_1. Then, players cascade "Whistle" messages and player $i-1$ plays P with a high probability.

The following theorem characterizes the condition where the RP strategy constitutes a belief-free equilibrium.

Theorem 3. *The sufficient and necessary condition in which the RP strategy constitutes a belief-free equilibrium is:*

$$\delta^{n-1}[1-(n+1)e]^{n-1} z \geq c. \tag{3}$$

Proof (Proof Sketch). Let X denote the following $n \times n$ matrix:

$$X = \begin{pmatrix} 1-(n-1)e & e & e & \dots & e & e \\ 1-(n-1)e & e & e & \dots & e & e \\ 2e & 1-ne & e & \dots & e & e \\ 2e & e & 1-ne & \dots & e & e \\ \vdots & \vdots & & \ddots & & \vdots \\ 2e & e & e & \dots & 1-ne & e \end{pmatrix}.$$

Let $\boldsymbol{\gamma}_{i,t}$ denote the probability distribution of the state of player i at period t, i.e., it is an n element vector $\begin{pmatrix} \gamma_{S_C} & \gamma_{S_P} & \gamma_{S_{W_{n-2}}} & \cdots & \gamma_{S_{W_1}} \end{pmatrix}$. Then, the probability distribution of the state of player $i+k$ at period $t+k$ is given as $\boldsymbol{\gamma}_{i,t} X^k$. Here, for all $k \geq n-1$, $X^k = X^{n-1}$ holds. By utilizing this fact, we can prove this theorem in a similar way as Theorem 1. □

6 Discussions

Additional Actions/Observations: Let us argue whether the new actions we introduced (i.e., W and P) are available in real-life situations. First, y, i.e., the decreased amount of player $i+1$'s benefit when player i plays W, can be 0. Thus, W can be basically identical to C, but it must be distinguished from C with a high probability. Such a new action (and an observation) would be easy to implement (e.g., by cheap talk [10]). On the other hand, z, i.e., the decreased amount of the benefit for player $i+1$ when player i plays P, must be large enough as shown in Theorem 1. However, within reasonable parameter settings, z can be smaller than 1. Thus, a relatively mild spiteful action, which decreases the value of the product, is sufficient.

Related Literature: In the literature of AI and multi-agent systems, there are many streams associated with repeated games [5]: the complexity of equilibrium computation [2, 4, 15], multi-agent learning [3, 6, 25], repeated congestion games [27], partially observable stochastic games (POSGs) [7, 13], and so on.

The repeated PD with imperfect observability has been extensively studied, but most papers assume public monitoring. A well-known result by Radner, Myerson, and Maskin [22] states that any pure strategy equilibrium payoff sum is bounded away from full efficiency however patient the players are. Abreu, Milgrom, and Pearce [1] explicitly derive an upper bound on the equilibrium total payoff. The literature on the repeated PD with imperfect private monitoring first studies sequential equilibria by randomized strategies under nearly-perfect monitoring (e.g., [23]), and then extends to arbitrarily noisy monitoring structures (e.g., [26]). By utilizing the concept of a belief-free equilibrium, analyzing equilibria becomes more tractable and what kinds of cooperative relations can be sustainable in the repeated PD has been examined [9, 8, 21].

To the best of our knowledge, we are the first to introduce the idea of the triangular trade in repeated games with imperfect private monitoring. Also, the idea of adding a dominated and seemingly irrelevant action is new, whether the monitoring is public or private. A notable exception is Shigenaka et al. [24], who show that by adding a dominated action, sustainable cooperation can be achieved in the repeated PD and in a problem called a team production problem. In this paper, we deal with a different problem, i.e., the triangular trade.

In evolutionary biology, several types of "reciprocity," i.e., mechanisms in which altruistic behavior may evolve, have been examined [19]. Among these works, our triangular trade resembles indirect reciprocity [20], in which a pair is randomly chosen, and one player acts as a donor while the other player acts as a recipient. If the donor chooses "Cooperate," she pays cost c and the recipient receives benefit b. If the donor chooses "Defect," both receive 0. It is shown that strategies based on reputation, i.e., helpful people are more likely to receive help, can evolve [20]. Our work is different from indirect reciprocity in the following points: (i) the same set of players repeatedly plays the game, and (ii) it deals with imperfect private monitoring and a belief-free equilibrium.

7 Conclusions and future works

In this paper, we proposed a new fundamental problem called triangular trade, which models a situation that is similar to the PD, i.e., the cooperation of players is efficient, but each player has an incentive to defect, while it is different from the PD since a player cannot directly punish a seemingly defected player. We first showed that when there exist only two actions, no simple strategy constitutes a belief-free equilibrium. Then, we showed that by adding two additional actions (and associated observations), the RP strategy can constitute a belief-free equilibrium in a wide range of parameter settings. Furthermore, we showed the RP strategy can be extended to the case with four or more players.

Our immediate future works include examining whether the expected reward of the RP strategy is theoretically optimal within all the strategies that constitute belief-free equilibria, as Shigenaka et al. [24] have shown for the two-player PD and team production problem.

Acknowledgement

This work was partially supported by JSPS KAKENHI Grant Number 17H00761 and JST, Strategic International Collaborative Research Program, SICORP.

References

1. Abreu, D., Milgrom, P., Pearce, D.: Information and timing in repeated partnerships. Econometrica 59, 1713–1733 (1991)
2. Andersen, G., Conitzer, V.: Fast equilibrium computation for infinitely repeated games. In: Proceedings of the 27th AAAI Conference on Artificial Intelligence (AAAI-13). pp. 53–59 (2013)
3. Blum, A., Monsour, Y.: Learning, regret minimization, and equilibria. In: Algorithmic game theory, pp. 79–101. Cambridge University Press (2007)
4. Borgs, C., Chayes, J., Immorlica, N., Kalai, A.T., Mirrokni, V., Papadimitriou, C.: The myth of the folk theorem. Games and Economic Behavior 70(1), 34 – 43 (2010)
5. Burkov, A., Chaib-draa, B.: Repeated games for multiagent systems: A survey. The Knowledge Engineering Review pp. 1–30 (2013)
6. Conitzer, V., Sandholm, T.: AWESOME: a general multiagent learning algorithm that converges in self-play and learns a best response against stationary opponents. Machine Learning 67(1), 23–43 (2007)
7. Doshi, P., Gmytrasiewicz, P.J.: On the Difficulty of Achieving Equilibrium in Interactive POMDPs. In: Proceedings of the 21st National Conference on Artificial Intelligence (AAAI-06). pp. 1131–1136 (2006)

8. Ely, J.C., Hörner, J., Olszewski, W.: Belief-free equilibria in repeated games. Econometrica 73(2), 377–415 (2005)
9. Ely, J.C., Välimäki, J.: A robust folk theorem for the prisoner's dilemma. Journal of Economic Theory 102(1), 84–105 (2002)
10. Farrell, J., Rabin, M.: Cheap talk. The Journal of Economic Perspectives 10(3), 103–118 (1996)
11. Fudenberg, D., Levine, D., Maskin, E.: The folk theorem with imperfect public information. Econometrica 62(5), 997–1039 (1994)
12. Fudenberg, D., Maskin, E.: The Folk Theorem in Repeated Games with Discounting or with Incomplete Information. Econometrica 54(3), 533–554 (1986)
13. Hansen, E.A., Bernstein, D.S., Zilberstein, S.: Dynamic programming for partially observable stochastic games. In: Proceedings of the 19th National Conference on Artificial Intelligence (AAAI-04). pp. 709–715 (2004)
14. Kreps, D.M., Wilson, R.: Sequential equilibria. Econometrica 50(4), 863–894 (1982)
15. Littman, M.L., Stone, P.: A polynomial-time Nash equilibrium algorithm for repeated games. Decision Support Systems 39(1), 55–66 (2005)
16. Maggi, G.: The role of multilateral institutions in international trade cooperation. The American Economic Review 89(1), 190–214 (1999)
17. Mailath, G.J., Samuelson, L.: Repeated Games and Reputations. Oxford University Press (2006)
18. Nowak, M., Sigmund, K.: A strategy of win-stay, lose-shift that outperforms tit-for-tat in prisoner's dilemma. Nature 364, 56–58 (1993)
19. Nowak, M.A.: Evolutionary dynamics. Harvard University Press (2006)
20. Nowak, M.A., Sigmund, K.: Evolution of indirect reciprocity by image scoring. Nature 393(6685), 573–577 (1998)
21. Piccione, M.: The repeated prisoner's dilemma with imperfect private monitoring. Journal of Economic Theory 102(1), 70–83 (2002)
22. Radner, R., Myerson, R., Maskin, E.: An example of a repeated partnership game with discounting and with uniformly inefficient equilibria. Review of Economic Studies 53, 59–69 (1986)
23. Sekiguchi, T.: Efficiency in repeated prisoner's dilemma with private monitoring. Journal of Economic Theory 76, 345–361 (1997)
24. Shigenaka, F., Sekiguchi, T., Iwasaki, A., Yokoo, M.: Achieving sustainable cooperation in generalized prisoner's dilemma with observation errors. In: Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI-17). pp. 677–683 (2017)
25. Shoham, Y., Leyton-Brown, K.: Learning and teaching. In: Multiagent systems: Algorithmic, Game-Theoretic, and Logical Foundations, pp. 189–222. Cambridge University Press (2008)
26. Sugaya, T.: Folk theorem in repeated games with private monitoring (2015), revised and resubmitted to Review of Economic Studies
27. Tennenholtz, M., Zohar, A.: Learning equilibria in repeated congestion games. In: Proceedings of the 8th International Joint Conference on Autonomous Agents and Multi-Agent System (AAMAS-09). pp. 233–240 (2009)