Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

Achieving Sustainable Cooperation in Generalized Prisoner's Dilemma with Observation Errors*

Fuuki Shigenaka,[1] **Tadashi Sekiguchi,**[2] **Atsushi Iwasaki,**[3] **Makoto Yokoo**[1]
1: Kyushu University, Motooka 744, Fukuoka, Japan. {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

A repeated game is a formal model for analyzing cooperation in long-term relationships, e.g., in the prisoner's dilemma. Although the case where each player observes her opponent's action with some observation errors (imperfect private monitoring) is difficult to analyze, a special type of an equilibrium called belief-free equilibrium is identified to make the analysis in private monitoring tractable. However, existing works using a belief-free equilibrium show that cooperative relations can be sustainable only in ideal situations.

We deal with a generic problem that can model both the prisoner's dilemma and the team production problem. We examine a situation with an additional action that is dominated by another action. To our surprise, by adding this seemingly irrelevant action, players can achieve sustainable cooperative relations far beyond the ideal situations. More specifically, we identify a class of strategies called one-shot punishment strategy that can constitute a belief-free equilibrium in a wide range of parameters. Moreover, for a two-player case, the obtained welfare matches a theoretical upper bound.

Introduction

A repeated game, where players repeatedly play the same stage game over an infinite time horizon, is a formal model for analyzing cooperation in long-term relationships and 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 repeated Prisoner's Dilemma (PD) game with action-errors, investigated by Nowak and Sigmund (1993). In contrast, suppose that each player observes her opponent's action with some observation errors. Assume that each player chooses cooperation (C) or defection (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 is more tractable (Ely and Välimäki 2002; Piccione 2002). Also, what kinds of cooperative relations can be sustainable in the repeated PD is examined (Ely, Hörner, and Olszewski 2005). However, these works show that cooperative relations can be sustainable only in ideal cases where the discount factor (δ) is close to 1 and/or the observation error rate (ϵ) is close to 0.

The original contributions of this paper are the following. We deal with a generic problem that can model both the repeated PD and the team production problem.[1] In the team production problem for two players, playing (C, D) or (D, C) maximizes players' total welfare, i.e., players should work (i.e., play C) and rest (i.e., play D) in turns. When there exist more than two players, players' total welfare is maximized when a certain number of players work and the other players rest.

Furthermore, we introduce an additional action that we call C' as well as an associated signal g' for this action. This action is dominated by another action, and playing it decreases the players' total welfare. Thus, adding it is irrelevant in a one-shot game, i.e., it does not affect equilibria at all; the probability that the action is taken is zero. As far as we are aware, we are the first to introduce such a dominated action and theoretically analyze equilibria in the generic game with imperfect private monitoring. Note that introducing such an action is not interesting with perfect or imperfect public monitoring, since it is well-known that co-

*This work was partially supported by JSPS KAKENHI Grant Number 24220003, 26280081, and 26380238.

[1]The works on the team production problem are scarce. Kobayashi, Ohta, and Sekiguchi (2014) study such a model, assuming public monitoring.

Table 1: Stage game payoff ($n = 2$, two actions)

	$a_2 = C$	$a_2 = D$
$a_1 = C$	$1, 1$	$-y, 1 + x$
$a_1 = D$	$1 + x, -y$	$0, 0$

operative relations are sustainable without introducing such an action due to the celebrated folk theorem (Fudenberg and Maskin 1986; Fudenberg, Levine, and Maskin 1994). 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 our newly introduced action does not change the players' minimax values.

To our surprise, it turns out that by adding this action, players can achieve sustainable cooperative relations far beyond the ideal cases identified in existing works. More specifically, we identify a class of strategies called one-shot punishment strategy that constitutes a belief-free equilibrium in a wide range of δ and ϵ. Moreover, when the number of players is two, we show that the sum of the discounted average payoffs achieved by the one-shot punishment strategies is actually theoretically optimal.

Here is the intuition why adding a seemingly irrelevant action is effective in the private monitoring case. In the repeated PD, a player needs to start punishment when she observes b to achieve sustainable cooperation; otherwise, her opponent clearly has an incentive to play D. However, if she uses D for punishment, her opponent might believe that she is defecting, and then the opponent starts punishment. Such an exchange of punishments will decrease their total discounted average payoffs and harm sustainable cooperation. When C' is available, players can use it for punishment. Since a player has no incentive to switch from C to C', C' and its associated signal g' can be used as a method to warn her opponent that she is starting punishment, which is less likely to cause an exchange of punishments due to a misunderstanding.

Preliminaries

Model

There exists a set of players $N = \{1, 2, \dots, n\}$. Each player $i \in N$ repeatedly plays the same stage game over an infinite horizon $t = 0, 1, 2, \dots$. In each period, player i takes some action a_i from a finite set A. Assume an action profile in that period is $\boldsymbol{a} = (a_1, \dots, a_n) \in A^n$. Then, her expected payoff in that period is given by stage game payoff function $u_i(\boldsymbol{a})$. Let $\boldsymbol{a}_{-i} = (a_1, \dots, a_{i-1}, a_{i+1}, \dots, a_n)$ denote the action profile of players $N \setminus \{i\}$.

Let us show an example for $N = \{1, 2\}$, $A = \{C, D\}$, where u is given as Table 1, where $x, y > 0$.

When $x < y + 1$, this stage game corresponds to the well-known PD, where (C, C) is the outcome that maximizes the total payoff of two players. When $x > y + 1$, this stage game corresponds to the team production problem (Kobayashi, Ohta, and Sekiguchi 2014), where the outcome that maximizes the total payoff of two players is either (C, D) or (D, C).

Within each period, player i observes her private signal $\omega_i^j \in \Omega$ that is related to player j's action. In the repeated PD with private monitoring, $\Omega = \{g, b\}$. Let $\vec{\omega}_i = (\omega_i^1, ..., \omega_i^{i-1}, \omega_i^{i+1}, ..., \omega_i^n) \in \Omega^{n-1}$ denote the profile of the private signals of player i. Also, let $o(\omega_i^j \mid a_j)$ denote the marginal distribution of ω_i^j given player j's action a_j. The signals are independent, i.e., the probability that player i receives the profile of private signals $\vec{\omega}_i$ when players take $\boldsymbol{a}_{-i}$ is given as $o(\vec{\omega}_i \mid \boldsymbol{a}_{-i}) = \prod_{j \in N \setminus \{i\}} o(\omega_i^j \mid a_j)$.

Let us describe a typical private monitoring scenario called *nearly-perfect* monitoring. When an opponent chooses C (or D), the "correct" signal is g (or b). A player receives a correct signal with high probability but she receives a wrong signal with small probability.

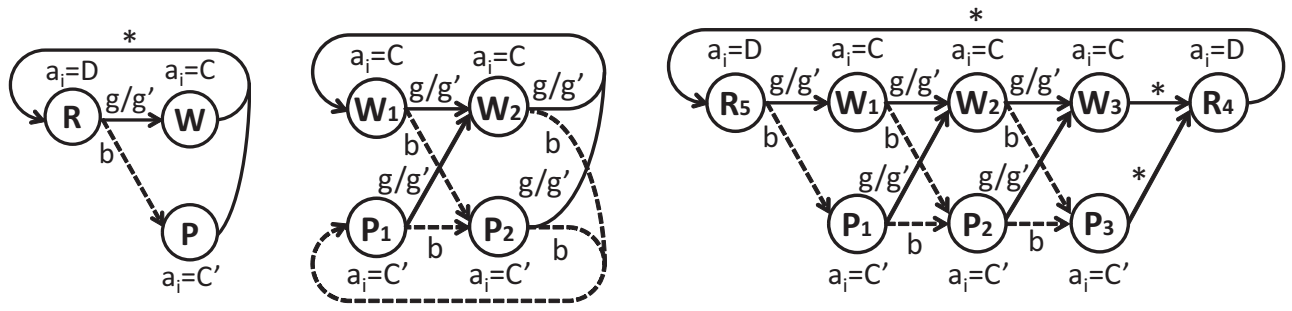
We assume no player can infer which action was taken (or not taken) by another player for sure: each signal $\omega_i^j \in \Omega$ occurs with a positive probability for any $a_j \in A$ (*full-support assumption*). Player i's *realized* payoff, which is determined by her own action and signals, is denoted as $\pi_i(a_i, \vec{\omega}_i)$. Hence, her expected payoff is given by $\sum_{\vec{\omega}_i \in \Omega^{n-1}} \pi_i(a_i, \vec{\omega}_i) \cdot o(\vec{\omega}_i \mid \boldsymbol{a}_{-i})$. We assume this expected value of the realized payoff is identical to stage game payoff $u_i(\boldsymbol{a})$. This formulation ensures that realized payoff π_i conveys no more information than a_i and $\vec{\omega}_i$. 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, we do not specify the particular values of the realized payoffs. This model is standard in the literature of repeated games with private monitoring (Mailath and Samuelson 2006).

Let us introduce a scenario called secret price-cutting, in which the assumption in this model, i.e., the realized payoff of a player is determined by her own action and signals, is appropriate. Assume that players are managers of two competing stores that have a price agreement. A manager can either keep the agreement (play C) or secretly cut her price (play D). A player's signal represents the number of customers who visit her store (g or b). The signal is affected by the action of the other player, but even when the other player keeps the agreement, the number of customers might be accidentally low. Also, her realized payoff is determined solely by her own action and signal, i.e., the price and the number of customers in her store.

Player i's expected discounted payoff from a sequence of action profiles $\boldsymbol{a}^0, \boldsymbol{a}^1, \dots$ 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 $(1 - \delta) \sum_{t=0}^{\infty} \delta^t u_i(\boldsymbol{a}^t)$.

Strategy representation and equilibrium concept

For player i, the set of her private histories at period t is $H_i^t := (A \times \Omega^{n-1})^t$. Each element $h_i^t = (a_i^0, \vec{\omega}_i^0, ..., a_i^{t-1}, \vec{\omega}_i^{t-1}) \in H_i^t$ 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 an ini-



(a) $n = 2, L^* = 1$ (b) $n = 2, L^* = 2$ (c) $n = 5, L^* = 3$

Figure 1: Pre-FSA of one-shot punishment strategy: $\sigma^{L^*/n}$

tial 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 (Kreps and Wilson 1982). 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 $T_i : \Theta_i \times \Omega^{n-1} \to \Theta_i$ specifies a deterministic state transition. Specifically, $T_i(\theta_i^t, \vec{\omega}_i^t)$ returns next state θ_i^{t+1} when the current state is θ_i^t and player i's private signal profile is $\vec{\omega}_i^t$. 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 $\boldsymbol{M}$ denote a profile of FSAs $(M_1, \ldots, M_n)$.

Figure 1(a) shows an example of a pre-FSA for a two-player stage game with three actions and three observations; $n = 2, A = \{C, D, C'\}$, and $\Omega = \{g, b, g'\}$. Here, a node represents a state and a directed link represents a state transition according to an observation. Since we assume $n = 2$, there is only one observation for each player. Symbol "$*$" represents any observation in $\{g, b, g'\}$. There are three states: W, R, and P. Each state represents a "working", "resting", or "punishing" state, and a player in each state should choose C, D, or C'.

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* $\boldsymbol{M} = ((m_1, \hat{\theta}_1), \ldots, (m_n, \hat{\theta}_n))$ *is a belief-free equilibrium if for all* t*, for all* $(\theta_1, \ldots, \theta_n) \in \prod_{i\in N} \Theta_i^t$*, 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.

Let us show an example. Let m denote the pre-FSA in Figure 1(a). Assume player 1 uses FSA (m, W) and player 2 uses (m, R), i.e., player 1 starts from state W and player 2 starts from state R at period 0. When t is an even number, $\Theta_1^t = \{W, P\}$ and $\Theta_2^t = \{R\}$. When t is an odd number, $\Theta_1^t = \{R\}$ and $\Theta_2^t = \{W, P\}$. In order that $((m, W), (m, R))$ is a belief-free equilibrium, we require the following: (i) (m, W) must be a best response when the other player is going to behave based on (m, R) and vice versa, and (ii) (m, P) must be a best response when the other player is going to behave based on (m, R) and vice versa. At each even period t, player 1 is either in state W or P, and player 2 is unsure where player 1 is located. Thus, she has a certain belief about this probability distribution. The fact that $((m, W), (m, R))$ is a belief-free equilibrium implies that behaving based on (m, R) is a best response for player 2, regardless of her belief. This is the reason that this equilibrium is called "belief-free", i.e., a player's best response is determined independently from her belief.

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 ($n = 2$, three actions)

	$a_2 = C$	$a_2 = D$	$a_2 = C'$
$a_1 = C$	1	$-y$	$1-\alpha$
$a_1 = D$	$1+x$	0	$1+x-\alpha$
$a_1 = C'$	1	$-y$	$1-\alpha$

Stage game

We introduce a generic game that can model both the repeated PD and the team production problem, where $A = \{C, D, C'\}$ and $\Omega = \{g, b, g'\}$. Here, C' is an additional action and g' is its "correct" signal. We set $o(\omega_i^j \mid a_j)$ to $1-2\epsilon$ when ω_i^j is the correct signal for a_j, and otherwise to ϵ. We assume $0 < \epsilon < 1/3$, i.e., a correct signal is more likely.

When $n = 2$, the stage game payoff is shown in Table 2. Here, we only show player 1's payoff since the game is symmetric. We assume $x, y > 0$. Here, the payoff for playing C' is identical to C for the player who plays it. On the other hand, the player can hurt the other player (by α) by playing C' instead of C. Since action C' is dominated by D, adding it is irrelevant when the stage game is played only once.

In general cases where $n \geq 2$, let $\#(A', \boldsymbol{a}_{-i})$ denote the number of actions in $\boldsymbol{a}_{-i}$, which are included in $A' \subseteq A$. We assume the stage game payoff is given as follows:

$$\begin{aligned} u_i((C, \boldsymbol{a}_{-i})) &= u_i((C', \boldsymbol{a}_{-i})) \\ &= v(\#(\{C, C'\}, \boldsymbol{a}_{-i})) - \alpha\#(\{C'\}, \boldsymbol{a}_{-i}), \\ u_i((D, \boldsymbol{a}_{-i})) &= v(\#(\{C, C'\}, \boldsymbol{a}_{-i})) - \alpha\#(\{C'\}, \boldsymbol{a}_{-i}) \\ &\quad + w(\#(\{C, C'\}, \boldsymbol{a}_{-i})). \end{aligned}$$

Here, v and w are some functions, which determine how many players should play C to maximize the total payoff. We assume $w(k)$ is non-negative for $0 \leq k \leq n-1$. When $n = 2$, if we set $v(0) = -w(0) = -y$, $v(1) = 1$ and $w(1) = x$, these payoffs become identical to those in Table 2.

In words, the payoff of player i is basically determined by the number of other players who play either C or C'. She incurs a loss in proportion to the number of other players who play C'. Also, if she plays D, she obtains a profit based on the number of other players who play either C or C'.

An action similar to C' would be available in actual application domains. For example, in the secret price-cutting scenario, C' can be considered a small price-cut. By doing so, the opponent suffers a certain loss, but the increase of her own customers is just enough to compensate the loss caused by the discount. Furthermore, let us assume there are two workers, one of whom is required to arrive at the office early in the morning and clean it. Doing so improves team productivity, but just one worker is adequate for cleaning. C' means that a worker shows up early but does not clean her co-worker's desk. Such a relatively mild spiteful action would be possible without incurring an additional cost.

One-shot punishment strategy

We define a class of pre-FSAs called *one-shot punishment* strategy. Different pre-FSA instances can be obtained by changing parameter L^* ($0 \leq L^* \leq n$), which represents the number of players who should play C simultaneously.

Two-player case

First, we consider the case with $n = 2$. When $L^* = 1$, players should work (i.e., play C) and rest (i.e., play D) in turns. When $L^* = 2$, both players should work at all times. Each player basically follows a prescribed cycle of two states as long as they observe "correct" signals. When $L^* = 1$, the action of one prescribed state is C, and the action of the other state is D. When $L^* = 2$, the actions of both states are C. Each player starts from different states in this prescribed cycle. When the game begins, each player monitors her opponent's behavior. Player i checks whether player j deviates to D when she should play C (or C'), while player i does not care about any other possible deviations (e.g., player j plays C instead of D). More precisely, if player i observes b when she is supposed to observe g (or g'), she punishes player j once by playing C' in the next period. If she detects a possible deviation by player j again (assuming player j continues the prescribed cycle), she punishes player j again. Otherwise, she returns to the prescribed cycle.

Figure 1(b) shows a pre-FSA where $L^* = 2$. Player 1 starts from W_1 and player 2 starts from W_2. Here, the upper-side cycle of W_1 and W_2 is the prescribed cycle. When player 1 is at W_1, player 2 should be at W_2 (or P_2) and should play C (or C'). If player 1 observes b, she punishes player 2 in the next period by moving to P_2. Figure 1(a) shows a pre-FSA where $L^* = 1$. Here, the prescribed cycle has two states: W and R. If player i is at W, player j is at R. Thus, player i never starts punishment in the next period. Therefore, only one punishment state P exists.

General case

Next, we consider cases with three or more players. Let $\sigma^{L^*/n}$ denote a pre-FSA for n players, in which L^* players are supposed to simultaneously work (i.e., play C). Figure 1(c) shows an example of a pre-FSA for $n = 5, L^* = 3$. As in the two-player case, each player basically follows a prescribed cycle of n states. First L^* states $W_1, \ldots, W_{L^*}$ are working states with prescribed action C, and the last $n - L^*$ states $R_{L^*+1}, \ldots, R_n$ are resting states with prescribed action D. Each player starts from her own initial state in this prescribed cycle. We assume player i starts from W_i when $1 \leq i \leq L^*$ and from R_i when $L^* + 1 \leq i \leq n$.

When the game begins, each player i monitors the behavior of player $i+1$ (player n monitors player 1). For example, player 1 monitors the behavior of player 2. Player i checks whether player $i+1$ deviates to D when she should play C (or C'), while player i does not care about other possible deviations (e.g., player $i+1$ plays C instead of D). More precisely, if player i observes b when her correct signal of player $i+1$ is supposed to be g (or g'), she punishes player $i+1$ once by playing C' in the next period.

Characteristics of one-shot punishment strategy

Let $E_N(\boldsymbol{s})$ denote the sum of all the players' discounted average payoffs obtained by strategy profile $\boldsymbol{s}$, i.e., $E_N(\boldsymbol{s}) = \sum_{i \in N} E_i(\boldsymbol{s})$. Let $\boldsymbol{\sigma}^{L^*/n}$ denote strategy profile

$((\sigma^{L^*/n}, W_1), \ldots, (\sigma^{L^*/n}, W_{L^*}), (\sigma^{L^*/n}, R_{L^*+1}), \ldots, (\sigma^{L^*/n}, R_n))$. The following theorem holds for $\boldsymbol{\sigma}^{L^*/n}$.

Theorem 1 *$\boldsymbol{\sigma}^{L^*/n}$ forms a belief-free equilibrium if and only if the following Inequality* (1) *holds:*

$$\alpha \geq \frac{w(L^*-1)}{\delta \cdot (1-3\epsilon)}. \tag{1}$$

Also, $E_N(\boldsymbol{\sigma}^{L^*/n})$ is given by the following Equation (2)*:*

$$\begin{aligned} &E_N(\boldsymbol{\sigma}^{L^*/n}) \\ &= L^* v(L^*-1) + (n-L^*)\{v(L^*)+w(L^*)\} \\ &\quad -(n-1)\delta\epsilon\alpha L^*. \end{aligned} \tag{2}$$

Proof. We are going to show that player i has no incentive to deviate from $\boldsymbol{\sigma}^{L^*/n}$, regardless of the current states of other players. Although there exist infinitely many possible deviations, by Proposition 12.2.3 in (Mailath and Samuelson 2006), it is sufficient to check the following three cases, each of which chooses a different action only once, and immediately returns to the original strategy[2]: (i) θ_i is W_k and player i chooses C' instead of C, or θ_i is P_k and player i chooses C instead of C', (ii) θ_i is W_k and player i chooses D instead of C, or θ_i is P_k and player i chooses D instead of C', (iii) θ_i is R_k and player i chooses C or C' instead of D.

There is only one player, j whose action is affected by player i's action (here, $j = i-1$ when $i \geq 2$, and $j = n$ when $i = 1$). Player i's action in current period t affects only player j's action in next period $t+1$, but it does not affect any other actions. Thus, it is sufficient to compare the sum of the discounted payoffs for two periods: periods t and $t+1$.

For Case (i), the payoff differences by taking this deviation are ± 0 in both periods; she has no incentive to do so.

For Case (ii), the differences are $+w(L^*-1)$ in period t and $-\alpha\delta\{o(b \mid D) - o(b \mid C)\}$ in period $t+1$. We require the sum must be at most 0. Then, we obtain Inequality (1).

For Case (iii), the differences are $-w(L^*)$ in period t and ± 0 in period $t+1$. Since the sum is negative, she has no incentive to choose this deviation.

From these results, we obtain that $\boldsymbol{\sigma}^{L^*/n}$ forms a belief-free equilibrium if and only if (1) holds.

Next, we calculate $E_N(\boldsymbol{\sigma}^{L^*/n})$. In each period, L^* players are either in W_k or P_k and choose C or C'. Then, $\#(\{C, C'\}, \boldsymbol{a}_{-i}) = L^* - 1$ for L^* players in W_k or P_k, while $\#(\{C, C'\}, \boldsymbol{a}_{-i}) = L^*$ for $n - L^*$ players in R_k.

When one player chooses C' instead of C, $\#(\{C'\}, \boldsymbol{a}_{-i})$ increases by one for the other players. Thus, if the number of players who choose C' instead of C increases by one, the sum of all the players' payoffs decreases by $(n-1)\alpha$. In the initial period, no player selects C', and $\#(\{C'\}, \boldsymbol{a}_{-i}) = 0$. For $t \geq 1$, each of the L^* players is either in W_k or P_k. The probability that she is actually in P_k equals ϵ, since the player monitored by her plays either C or C', and $o(b \mid C) = o(b \mid C') = \epsilon$. Thus, the effect that several players choose C' instead of C toward the total discounted average payoff is given as: $-(n-1)\delta\epsilon\alpha L^*$. □

[2]This property is called one-shot deviation principle or one-deviation property (Mailath and Samuelson 2006).

One possible implication of Theorem 1 is that, for any given δ, x, y, and ϵ, if we appropriately set the amount of punishment α, $\boldsymbol{\sigma}^{L^*/n}$ forms a belief-free equilibrium. Let us show several concrete examples. Strategy $\boldsymbol{\sigma}^{1/2}$ forms a belief-free equilibrium when $\alpha \geq y/\delta(1-3\epsilon)$ holds. The sum of the players' discounted average payoffs is given as $E_N(\boldsymbol{\sigma}^{1/2}) = -y + (1+x) - \delta\epsilon\alpha$. Let $E_N^*(\boldsymbol{\sigma}^{L^*/n})$ denote the maximum of $E_N(\boldsymbol{\sigma}^{L^*/n})$ by varying α in the range where Inequality (1) is satisfied. Then $E_N^*(\boldsymbol{\sigma}^{1/2})$ is given as:

$$E_N^*(\boldsymbol{\sigma}^{1/2}) = -y + (1+x) - \frac{\epsilon \cdot y}{1-3\epsilon}.$$

If $\delta = 0.9$, $\epsilon = 0.05$, and $y = x = 1$, the optimal/minimum value of α is around 1.3. Since $u_i((C, C')) \approx -0.3$, the degree of punishment is large enough to offset the mutual cooperation gain. On the other hand, since $u_i((D, C')) \approx 0.7$, it is not large enough to offset the gain of deviation.

Similarly, $\boldsymbol{\sigma}^{2/2}$ forms a belief-free equilibrium when $\alpha \geq x/\delta(1-3\epsilon)$ holds. Then $E_N^*(\boldsymbol{\sigma}^{2/2})$ is given as follows:

$$E_N^*(\boldsymbol{\sigma}^{2/2}) = 2 - 2\frac{\epsilon \cdot x}{1-3\epsilon}.$$

Furthermore, $\boldsymbol{\sigma}^{0/2}$ is a belief-free equilibrium for any $\alpha \geq 0$. Then, $E_N^*(\boldsymbol{\sigma}^{0/2})$ is given as follows: $E_N^*(\boldsymbol{\sigma}^{0/2}) = 0$. Note that $\boldsymbol{a} = (D, \ldots, D)$ is a Nash equilibrium of the stage game for any $\alpha \geq 0$.

We show that these values are actually *optimal* for any belief-free equilibria when $n = 2$ by the following theorem.

Theorem 2 *When $n = 2$, for any given δ, x, y, and ϵ, if a profile of strategies $\boldsymbol{s}$ forms a belief-free equilibrium, then the following Inequality* (3) *holds:*

$$E_N(\boldsymbol{s}) \leq \max(E_N^*(\boldsymbol{\sigma}^{2/2}), E_N^*(\boldsymbol{\sigma}^{1/2}), E_N^*(\boldsymbol{\sigma}^{0/2})). \tag{3}$$

Proof. Assume $\boldsymbol{s}$ gives the maximum E_N in strategies that constitute belief-free equilibria. Let $E_i(\boldsymbol{s} \mid \omega_i, \omega_j)$ denote the discounted average payoff of player i after she observes ω_i and her opponent j observes ω_j in period 1 when their strategy profile is $\boldsymbol{s}$. Since $\boldsymbol{s}$ forms a belief-free equilibrium, this continuation payoff depends only on ω_j, i.e., the following condition holds:

$$\forall \omega_i, \omega_i', \omega_j \in \Omega, E_i(\boldsymbol{s} \mid \omega_i, \omega_j) = E_i(\boldsymbol{s} \mid \omega_i', \omega_j).$$

Let $E_i(\boldsymbol{s} \mid \omega_j)$ denote $E_i(\boldsymbol{s} \mid \omega_i, \omega_j)$ for any $\omega_i \in \Omega$. We can rewrite $E_i(\boldsymbol{s})$ as follows:

$$(1-\delta)u_i(\boldsymbol{a}) + \delta \sum_{\omega_j \in \Omega} o(\omega_j \mid a_i) E_i(\boldsymbol{s} \mid \omega_j). \tag{4}$$

Since $\boldsymbol{s}$ provides the maximum E_N in all equilibrium strategies, the following condition is satisfied[3]:

$$\forall \omega_1, \omega_2 \in \Omega, E_N(\boldsymbol{s}) \geq E_1(\boldsymbol{s} \mid \omega_2) + E_2(\boldsymbol{s} \mid \omega_1). \tag{5}$$

In addition, player i has no incentive to change her action from a_i to a_i' if the following condition is satisfied:

$$\begin{aligned} &(1-\delta)\{u_i(a_i, a_j) - u_i(a_i', a_j)\} \\ &+\delta \sum_{\omega_j \in \Omega} \{o(\omega_j \mid a_i) - o(\omega_j \mid a_i')\} E_i(\boldsymbol{s} \mid \omega_j) \geq 0. \end{aligned}$$

[3]If this condition does not hold, $\boldsymbol{s}$ can be modified so that $E_N(\boldsymbol{s})$ is improved, while it still forms a belief-free equilibrium.

First, let us assume the prescribed action profile by $\boldsymbol{s}$ in the initial period is (C, a_j). The fact that player i has no incentive to change her action to C' or D is given as:

$$E_i(\boldsymbol{s} \mid g') \leq E_i(\boldsymbol{s} \mid g),$$

$$E_i(\boldsymbol{s} \mid b) \leq E_i(\boldsymbol{s} \mid g) - \frac{(1-\delta)\{u_i(D, a_j) - u_i(C, a_j)\}}{\delta(1-3\epsilon)}.$$

From the above, we obtain:

$$\sum_{\omega_j \in \{g, g', b\}} o(\omega_j \mid C) E_i(\boldsymbol{s} \mid \omega_j) \leq E_i(\boldsymbol{s} \mid g) - \epsilon \frac{(1-\delta)\{u_i(D, a_j) - u_i(C, a_j)\}}{\delta(1-3\epsilon)}. \quad (6)$$

When $a_j = C$, from Inequalities (5) and (6), we obtain:

$$\begin{aligned} & \sum_{\omega_2 \in \Omega} o(\omega_2 \mid C) E_1(\boldsymbol{s} \mid \omega_2) + \sum_{\omega_1 \in \Omega} o(\omega_1 \mid C) E_2(\boldsymbol{s} \mid \omega_1) \\ \leq \quad & E_1(\boldsymbol{s} \mid g) + E_2(\boldsymbol{s} \mid g) - 2\epsilon \frac{(1-\delta)x}{\delta(1-3\epsilon)} \\ \leq \quad & E_N(\boldsymbol{s}) - 2\epsilon \frac{(1-\delta)x}{\delta(1-3\epsilon)}. \end{aligned}$$

From the above and Equation (4), we obtain:

$$E_N(\boldsymbol{s}) \leq (1-\delta)2 + \delta\Big\{E_N(\boldsymbol{s}) - 2\epsilon \frac{(1-\delta)x}{\delta(1-3\epsilon)}\Big\}.$$

Therefore, we obtain:

$$E_N(\boldsymbol{s}) \leq 2 - 2\frac{\epsilon \cdot x}{1-3\epsilon} = E_N^*(\boldsymbol{\sigma}^{2/2}).$$

Similarly, when $a_j = D$, we obtain:

$$E_N(\boldsymbol{s}) \leq (1 + x - y) - \frac{\epsilon \cdot y}{1-3\epsilon} = E_N^*(\boldsymbol{\sigma}^{1/2}).$$

Next, we assume the prescribed action profile by $\boldsymbol{s}$ in period 0 is (D, D). From condition (5), we obtain:

$$E_N(\boldsymbol{s}) \leq (1-\delta)0 + \delta E_N(\boldsymbol{s}).$$

Then, we obtain $E_N(\boldsymbol{s}) \leq 0 = E_N^*(\boldsymbol{\sigma}^{0/2})$.

We need to check similar conditions where the prescribed action profile in period 0 is (C', C'), (C', D), or (C, C') but we only obtain weaker conditions. Thus, from the above discussions, we obtain Inequality (3). □

Discussions

Nondeterministic strategy: For notational simplicity, we describe a strategy with pure actions and deterministic state transitions. However, our analysis is applicable to any general strategy that includes mixed action and nondeterministic state transitions, since its expected utility is just a linear combination of values obtained in the proof. Thus, our one-shot punishment strategy is optimal even in all general strategies including nondeterministic strategies.

Welfare for $n \geq 3$: When three or more players exist, we cannot prove that the welfare obtained by a profile of one-shot punishment strategies is optimal. In a one-shot punishment strategy, the action of player $i+1$ is only monitored by player i. Thus, players might be able to integrate their observations and to detect possible deviations more precisely. However, such a sophisticated strategy seems very complex; we currently have no good idea for constructing it.

Related literature: In the literature of AI and multi-agent systems, there are many streams associated with repeated games (Burkov and Chaib-draa 2013): the complexity of equilibrium computation (Andersen and Conitzer 2013; Borgs et al. 2010; Littman and Stone 2005), multi-agent learning (Blum and Monsour 2007; Conitzer and Sandholm 2007; Shoham and Leyton-Brown 2008), repeated congestion games (Tennenholtz and Zohar 2009), partially observable stochastic games (POSGs) (Doshi and Gmytrasiewicz 2006; Hansen, Bernstein, and Zilberstein 2004), 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 (1986) states that any pure strategy equilibrium payoff sum is bounded away from full efficiency however patient the players are. Abreu, Milgrom, and Pearce (1991) extend the analysis and explicitly derive an upper bound on the equilibrium total payoff. The upper bound obtained in this paper has a similar structure. One contribution of our work is generalizing their results to private monitoring settings.

The literature on the repeated PD with imperfect private monitoring first studies sequential equilibria by randomized strategies under nearly-perfect monitoring (e.g., Sekiguchi 1997), and then extends to arbitrarily noisy monitoring structure (e.g., Sugaya 2015). This paper also allows arbitrarily noisy monitoring and adopts the belief-free equilibrium (Ely, Hörner, and Olszewski 2005; Ely and Välimäki 2002) by pure strategies. Piccione (2002) constructs cooperative equilibria in the repeated PD under ideal situations based on Sekiguchi (1997). Because of the third action, we are able to provide an upper bound on the total equilibrium payoffs and construct simple belief-free equilibria that attain the bound even if the players are not excessively patient.

Conclusions

This paper investigated repeated games with imperfect private monitoring. We introduced a generic problem that can model both the repeated PD and the team production problem, and examined a situation where seemingly irrelevant (i.e., dominated) action C' is added. We identified the one-shot punishment strategy, which can constitute a belief-free equilibrium in a wide range of parameters. The strategy is easy to adapt since it is a pure strategy and is concisely represented by an FSA. Moreover, when the number of players is two, we showed that the obtained welfare of this equilibrium matches a theoretical upper bound. Our future works will apply a similar idea to different settings, e.g., C' requires an additional cost, a player can choose the level of punishment, and so on.

References

Abreu, D.; Milgrom, P.; and Pearce, D. 1991. Information and timing in repeated partnerships. *Econometrica* 59:1713–1733.

Andersen, G., and Conitzer, V. 2013. Fast equilibrium computation for infinitely repeated games. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence (AAAI-13)*, 53–59.

Blum, A., and Monsour, Y. 2007. Learning, regret minimization, and equilibria. In *Algorithmic game theory*. Cambridge University Press. 79–101.

Borgs, C.; Chayes, J.; Immorlica, N.; Kalai, A. T.; Mirrokni, V.; and Papadimitriou, C. 2010. The myth of the folk theorem. *Games and Economic Behavior* 70(1):34 – 43.

Burkov, A., and Chaib-draa, B. 2013. Repeated games for multiagent systems: A survey. *The Knowledge Engineering Review* 1–30.

Conitzer, V., and Sandholm, T. 2007. 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.

Doshi, P., and Gmytrasiewicz, P. J. 2006. On the Difficulty of Achieving Equilibrium in Interactive POMDPs. In *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI-06)*, 1131–1136.

Ely, J. C., and Välimäki, J. 2002. A robust folk theorem for the prisoner's dilemma. *Journal of Economic Theory* 102(1):84–105.

Ely, J. C.; Hörner, J.; and Olszewski, W. 2005. Belief-free equilibria in repeated games. *Econometrica* 73(2):377–415.

Fudenberg, D., and Maskin, E. 1986. The Folk Theorem in Repeated Games with Discounting or with Incomplete Information. *Econometrica* 54(3):533–54.

Fudenberg, D.; Levine, D.; and Maskin, E. 1994. The folk theorem with imperfect public information. *Econometrica* 62(5):997–1039.

Hansen, E. A.; Bernstein, D. S.; and Zilberstein, S. 2004. Dynamic programming for partially observable stochastic games. In *Proceedings of the 19th National Conference on Artificial Intelligence (AAAI-04)*, 709–715.

Kobayashi, H.; Ohta, K.; and Sekiguchi, T. 2014. Repeated partnerships with decreasing returns. Public Economics Seminar, Keio University.

Kreps, D. M., and Wilson, R. 1982. Sequential equilibria. *Econometrica* 50(4):863–894.

Littman, M. L., and Stone, P. 2005. A polynomial-time Nash equilibrium algorithm for repeated games. *Decision Support Systems* 39(1):55–66.

Mailath, G. J., and Samuelson, L. 2006. *Repeated Games and Reputations*. Oxford University Press.

Nowak, M., and Sigmund, K. 1993. A strategy of win-stay, lose-shift that outperforms tit-for-tat in prisoner's dilemma. *Nature* 364:56–58.

Piccione, M. 2002. The repeated prisoner's dilemma with imperfect private monitoring. *Journal of Economic Theory* 102(1):70–83.

Radner, R.; Myerson, R.; and Maskin, E. 1986. An example of a repeated partnership game with discounting and with uniformly inefficient equilibria. *Review of Economic Studies* 53:59–69.

Sekiguchi, T. 1997. Efficiency in repeated prisoner's dilemma with private monitoring. *Journal of Economic Theory* 76:345–361.

Shoham, Y., and Leyton-Brown, K. 2008. Learning and teaching. In *Multiagent systems: Algorithmic, Game-Theoretic, and Logical Foundations*. Cambridge University Press. 189–222.

Sugaya, T. 2015. Folk theorem in repeated games with private monitoring. Revised and resubmitted to Review of Economic Studies.

Tennenholtz, M., and Zohar, A. 2009. Learning equilibria in repeated congestion games. In *Proceedings of the 8th International Joint Conference on Autonomous Agents and Multi-Agent System (AAMAS-09)*, 233–240.