Coordination via Mutual Punishment: POMDP Approach to Repeated Prisoner's Dilemma with Private Monitoring

Atsushi Iwasaki[*], YongJoon Joe[*], Michihiro Kandori[†], Ichiro Obara[‡], and Makoto Yokoo[*]

May 12, 2011

Abstract

The present paper identifies equilibria in an infinitely repeated prisoner's dilemma game, where each player privately receives a noisy signal about the opponent's action. This is an instance of repeated games with *private monitoring*, which have been paid considerable attention in the AI and economics literature. Finding pure strategy equilibria in such a game has been considered to be extremely hard. Recently, Kandori and Obara (2010) showed that the theory of the partially observable Markov decision process can be used to identify equilibria, when equilibrium behavior is described by a finite state automaton (FSA). Building upon this idea, we first examine an FSA that follows a well-known equilibrium strategy under perfect monitoring, grim-trigger (GT). Under GT, a player first cooperates, but as soon as she observes defection, she defects forever. We identify the range of signal parameters over which GT is an equilibrium. This is a region where the probability of errors is relatively small.

A shortcoming of GT is that it is too unforgiving and thus generates a low payoff. To cope with this problem, we propose a class of FSA called k-period mutual punishment (k-MP). Under this FSA, a player first cooperates. If she observes defection, she also defects, but after k consecutive periods of mutual defection, she returns to cooperation. We show that introducing such periods of mutual defection is beneficial in establishing a robust coordination among

[*]Kyushu University, Japan, {iwasaki@, yongjoon@agent., yokoo@}inf.kyushu-u.ac.jp
[†]University of Tokyo, Japan, kandori@e.u-tokyo.ac.jp
[‡]UCLA, California, ichiro.obara@gmail.com

players, i.e., the actions of players are well organized/synchronized after a deviation. A parameter k represents the degree of forgiveness; when $k = 1$, it is equivalent to a very forgiving strategy known as Pavlov, and when $k = \infty$, it is equivalent to GT. We show that 2-MP can be an equilibrium in a reasonably wide range of signal parameters although the size of the range is smaller than GT, and that its efficiency is significantly higher than GT. We exhaustively search for simple FSAs (with at most three states) and confirm that there exists no other FSA that constitutes an equilibrium in a reasonably wide range of signal parameters and is more efficient than GT.

1 Introduction

We consider repeated games with *imperfect private monitoring*, where each player privately receives a noisy observation (signal) of the opponent's action. This class of games represents long-term relationship among players and has a wide range of applications. Therefore, it has been paid considerable attention in the AI and economics literature. In particular, for the AI community, the framework has become increasingly important for handling noisy environments. In fact, Wang, Chatterjee, and Kwiat (2009) investigate strategies in ad hoc networks with noisy channels. Tennenholtz and Zohar (2009) consider repeated congestion games where an agent has limited capability in monitoring the actions of her counterparts.

Although a vast number of simulation-based studies have been conducted, analytical studies on this class of games have not been quite successful. This is because finding pure strategies equilibria in such games has been considered to be extremely hard, since it requires very complicated statistical inferences [Kandori, 2010]. As a result, such an analysis has been possible for only very restricted distributions of signals.

Quite recently, Kandori and Obara (2010) show that the theory of the partially observable Markov decision process (POMDP) can be used to identify equilibria, when equilibrium behavior is described by a finite state automaton (FSA). They further provide a tractable computational method to determine whether a given profile of finite state automata can constitute an equilibrium. Accordingly, this enables us to analyze equilibria for arbitrary distributions of signals.

Building upon the framework of Kandori and Obara (2010), we first examine an FSA that follows a well-known equilibrium strategy under perfect monitoring called grim-trigger (GT, Fig. 1).[1] We identify the range of signal parameters over

[1]Here, g or b is a private signal (*good* or *bad*), which is a noisy observation of the opponent's

which GT is an equilibrium. This is a region where the probability of errors is relatively small.

A shortcoming of GT is that it is too unforgiving and thus generates a low payoff. A well-known, more forgiving strategy is tit-for-tat (TFT, Fig. 2). However, if two players use TFT, an observation of defection leads to poorly coordinated behavior. The players alternate between (C, D) and (D, C) with a large probability. They can return to mutual cooperation only when both have observed cooperation in the previous period, which is very unlikely to happen.

To cope with this problem, we propose a novel FSA called k-period mutual punishment (k-MP). Figure 7 shows the FSA of 2-MP. Under this FSA, a player first cooperates. If her opponent defects, she also defects, but after two consecutive periods of mutual defection, she returns to cooperation. We can control the forgiveness of k-MP by changing the parameter k. Note that k-MP incorporates GT and a well-known strategy Pavlov [Kraines and Kraines, 1989] as a special case ($k = \infty$ or $k = 1$). The behavior of k-MP might look somewhat counterintuitive: A player returns to cooperation after observing a *bad* signal. However, requiring such periods of mutual defection is beneficial in establishing a robust coordination among players. We show that 2-MP can be an equilibrium in a reasonably wide range of signal parameters and that its efficiency is significantly higher than GT. Furthermore, we exhaustively search for simple FSAs (with at most three states) and confirm that no other FSA can constitute an equilibrium in a reasonably wide range of signal parameters and is more efficient than GT.

Traditionally, the POMDP framework is a popular approach for single-agent planning, and game theory has been extensively used for analyzing multi-agent interactions. However, there exist very few works on connecting these two areas; as mentioned in a popular AI textbook, "... there is currently no good way to combine game theoretic and POMDP control strategies" [Russell and Norvig, 2002]. One notable exception is Doshi and Gmytrasiewicz (2006). They investigate the computational complexity of deriving another class of equilibria. In their model, a player is more sophisticated than an FSA. As a result, their equilibrium is more involved and reaching such an equilibrium is implausible in practice. The authors believe this paper narrows the gap between game theory and POMDP literature and introduces a new exciting interdisciplinary research field between these areas.

---

action C or D.

2 Repeated games with private monitoring

We model a repeated game with private monitoring according to [Kandori and Obara, 2010]. We concentrate on two-player, symmetric games (where a game is invariant under the permutation of players' identifiers). However, the techniques introduced in this paper can be easily extended to n-player, non-symmetric cases.

Player $i \in \{1, 2\}$ repeatedly plays the same stage game over an infinite horizon $t = 1, 2, \ldots$. In each period, player i takes some action a_i from a finite set A, and her expected payoff in that period is given by a stage game payoff function $g_i(\boldsymbol{a})$, where $\boldsymbol{a} = (a_1, a_2) \in A^2$ is the action profile in that period. Within each period, player i observes her private signal $\omega_i \in \Omega$. Let $\boldsymbol{\omega}$ denote an observation profile $(\omega_1, \omega_2) \in \Omega^2$ and let $o(\boldsymbol{\omega} \mid \boldsymbol{a})$ be the probability of private signal profile $\boldsymbol{\omega}$ given an action profile $\boldsymbol{a}$. We assume that Ω is a finite set and we denote the marginal distribution of ω_i by $o_i(\omega_i \mid \boldsymbol{a})$. It is also assumed that no player can infer which action was taken (or not taken) by another player for sure; to this end, we assume that each signal profile $\boldsymbol{\omega} \in \Omega^2$ occurs with a positive probability for any $\boldsymbol{a} \in A^2$.

Player i's *realized* payoff is determined by her own action and signal and denoted $\pi_i(a_i, \omega_i)$. Hence, her expected payoff is given by $g_i(\boldsymbol{a}) = \sum_{\boldsymbol{\omega} \in \Omega^2} \pi_i(a_i, \omega_i) o(\boldsymbol{\omega} \mid \boldsymbol{a})$. This formulation ensures that the realized payoff π_i conveys no more information than a_i and ω_i do. Note that the expected payoff is determined by the action profile, while the realized payoff is determined solely by her own action and signal.

Let us motivate this model by an example. Assume players are managers of two competing stores. The action of each player is to determine the price of an item in her store. The signal of a player represents the number of customers who visit her store. The signal is affected by the action of another player, i.e., the price of the competing store, but the realized payoff is determined solely by her own action and signal, i.e., the price and the number of customers.

The stage game is to be played repeatedly over an infinite time horizon. Player i's discounted payoff G_i from a sequence of action profiles $\boldsymbol{a}^1, \boldsymbol{a}^2, \ldots$ is $\sum_{t=1}^{\infty} \delta^t g_i(\boldsymbol{a}^t)$, with discount factor $\delta \in (0, 1)$. Also, the discounted *average payoff* (payoff per period) is defined as $(1 - \delta) G_i$.

2.1 Repeated game strategies and finite state automata

We introduce some basic terms and notations. A private history for player i at the end of time t is the record of player i's past actions and signals, $h_i^t = (a_i^0, \omega_i^0, \ldots, a_i^t, \omega_i^t) \in H_i^t := (A \times \Omega)^{t+1}$. To determine the initial action of each

player, we introduce a dummy initial history (or null history), h_i^0 and let H_i^0 be a singleton set $\{h_i^0\}$. A pure strategy s_i for player i is a function specifying an action after any history, or, formally, $s_i : H_i \to A$, where $H_i = \bigcup_{t \geq 0} H_i^t$.

A finite state automaton (FSA) is a popular approach for compactly representing the behavior of a player. An FSA M is defined by $\langle \Theta, \hat{\theta}, f, T \rangle$, where Θ is a set of states, $\hat{\theta} \in \Theta$ is an initial state, $f : \Theta \to A$ determines the action choice for each state, and $T : \Theta \times \Omega \to \Theta$ specifies a deterministic state transition. Specifically, $T(\theta^t, \omega^t)$ returns the next state θ^{t+1} when the current state is θ^t and the private signal is ω^t. We call an FSA without the specification of the initial state, i.e., $m = \langle \Theta, f, T \rangle$, as a finite state *preautomaton* (pre-FSA). Now, we introduce a *symmetric pure finite state equilibrium*.

Definition 1. *A symmetric pure finite state equilibrium (SPFSE) is a pure strategy sequential equilibrium of a repeated game with private monitoring, where each player's behavior on the equilibrium path is given by an FSA $M = \langle \Theta, \hat{\theta}, f, T \rangle$.*

A sequential equilibrium is a refinement of Nash equilibrium for games of imperfect information. In this definition, we require that an FSA specifies only the behavior of a player on the equilibrium path (please consult [Kandori and Obara, 2010] for detail). We briefly describe this point later.

It must be emphasized that if an FSA M constitutes an equilibrium, it means that as long as player 2 behaves according to M, player 1's *best response* is to act according to M. Here, we do not restrict the possible strategy space of player 1 at all. More specifically, M is the best response not only within strategies that can be represented as FSAs but also within all possible strategies, including strategies that require an infinite number of states.

2.2 Private monitoring problem as POMDP

We give a procedure to check whether an FSA $M = \langle \Theta, \hat{\theta}, f, T \rangle$ constitutes an SPFSE. First, by assuming each player acts according to an FSA M, we can create a joint FSA. The expected discounted payoff of this joint FSA for player 1 is given as $V_{\hat{\theta},\hat{\theta}}$, where V_{θ_1,θ_2} can be obtained by solving a system of linear equations defined as follows.

$$V_{\theta_1,\theta_2} = g_1((f(\theta_1), f(\theta_2))) \\ +\delta \sum_{(\omega_1,\omega_2) \in \Omega^2} o((\omega_1, \omega_2) \mid (f(\theta_1), f(\theta_2))) \cdot V_{T(\theta_1,\omega_1),T(\theta_2,\omega_2)}.$$

Now, let us consider how to obtain a best response for player 1, assuming player 2 behaves according to M. Player 1 confronts a Markov decision process,

where the state of the world is represented by the state of player 2's FSA. However, player 1 cannot directly observe the state of player 2. Thus, this problem is equivalent to finding an optimal policy in POMDP.

More precisely, the POMDP of this problem is defined by $\langle \Theta, A, \Omega, O, P, R \rangle$, where Θ is a set of states of player 2, A is a set of actions of player 1, Ω is a set of observations of player 1, O represents an observation probability function, P represents a state transition function, and R is a payoff function. Θ, A, and Ω are already defined. $O(\omega_1 \mid a_1, \theta^t)$ represents the conditional probability of observing ω_1 after performing an action a_1 at a state θ^t (of player 2), which is defined as: $O(\omega_1 \mid a_1, \theta^t) = o_1(\omega_1 \mid (a_1, f(\theta^t)))$.

In a standard POMDP model, we usually assume the observation probability depends on the next state θ^{t+1} rather than on the current state θ^t. We use this slightly different model since it is more suitable for representing repeated games with private monitoring. However, this difference is insignificant and we can use the models interchangeably.

$P(\theta^{t+1} \mid \theta^t, a_1)$ represents the conditional probability that the next state is θ^{t+1} when the current state is θ^t and the action of player 1 is a_1, which is defined as:

$$P(\theta^{t+1} \mid \theta^t, a_1) = \sum_{\omega_2 \in \Omega | T(\theta^t, \omega_2) = \theta^{t+1}} o_2(\omega_2 \mid (a_1, f(\theta^t))).$$

An expected payoff function $R : A, S \to \mathbb{R}$ is given as: $R(a_1, \theta^t) = g_1((a_1, f(\theta^t)))$.

We can check whether an FSA $M = \langle \Theta, \hat{\theta}, f, T \rangle$ constitutes an SPFSE by using the following procedure. This procedure is based on the idea presented in [Kandori and Obara, 2010], but our description is more concrete and clearly specifies a way of utilizing an existing POMDP solver.

1. First solve a system of linear equations of a joint FSA and obtain the expected discounted payoff of player 1, i.e., $V_{\hat{\theta},\hat{\theta}}$, when both players follow M.
2. Obtain an optimal policy Π^* (which is given as a pre-FSA) and its value function $v(\cdot)$ for the POMDP $\langle \Theta, A, \Omega, O, P, R \rangle$. We can use a standard POMDP solver such as [Kaelbling, Littman, and Cassandra, 1998]. In general, this computation might not converge and no optimal policy can be represented as a pre-FSA. In such a case, we terminate the computation and obtain a semi-optimal policy.
3. Let us denote the belief of player 1 such that player 2 is in $\hat{\theta}$ for sure, as $b_{\hat{\theta}}$. If $v(b_{\hat{\theta}}) = V_{\hat{\theta},\hat{\theta}}$, then the FSA $M = \langle \Theta, \hat{\theta}, f, T \rangle$ constitutes an SPFSE.

To be more precise, due to the cancellation of the significant digit, checking whether $v(b_{\hat{\theta}}) = V_{\hat{\theta},\hat{\theta}}$ holds can be difficult. To avoid this problem, we need to

check the obtained optimal policy Π^* as well. Note that even if Π^* is not exactly the same as a pre-FSA m of M, the FSM can constitute an SPFSE. This is because, there can be a belief state that is unreachable when players act according to M. m does not need to specify the optimal behavior in such a belief state, while Π^* does specify the optimal behavior for all possible belief states.

To verify whether M constitutes an SPFSE, we first find the initial state θ^* in Π^* that is optimal when the other player employs M. Next, we examine a part of Π^*, i.e., the states that are reachable from θ^*, and check whether this part is coincident with M. Then, M is a best response to itself and thus it constitutes an SPFSE. In general, there can be multiple optimal policies and a POMDP solver usually returns just one optimal policy. Thus, even if one optimal policy Π^* does not include m, there might be another optimal policy that does include m. However, we can use m as an initial policy, and make sure that Π^* includes m as long as M constitutes an SPFSE.

3 Repeated prisoner's dilemma with noisy observation

We apply the POMDP technique to the prisoner's dilemma model analyzed by [Kandori and Obara, 2010]. The stage game payoff is given by

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

Each player's private signal is $\omega_i \in \{g, b\}$ (*good* or *bad*), which is a noisy observation of the opponent's action. For example, when the opponent chooses C, player i is more likely to receive the correct signal $\omega_i = g$, but sometimes an observation error provides a wrong signal $\omega_i = b$. Precisely, we assume that exactly one player receives a wrong signal with probability $q > 0$ and both receive wrong signals with $r > 0$. When the action profile is (C, C), the joint distribution of private signals $o(\boldsymbol{\omega} \mid \boldsymbol{a})$ is given as follows (when the action profile is (D, D), p and r are exchanged).

	$w_2 = g$	$w_2 = b$
$w_1 = g$	p	q
$w_1 = b$	q	r

Similarly, when the action profile is (C, D), the joint distribution of private signals is given as follows (when the action profile is (D, C), p and r are exchanged).

	$w_2 = g$	$w_2 = b$
$w_1 = g$	q	r
$w_1 = b$	p	q

We assume that $p \in (1/2, 1)$ and $q \in (0, 1/4)$ under the constraint $p+2q+r = 1$ and that $\pi_i(a_i, \omega_i)$ is chosen so that $g_i(\boldsymbol{a})$ is constant. This signal distribution naturally represents a noisy signal about the opponent's action. Throughout our paper, we basically use the default setting: $x = 1$, $y = 1$, and the discount factor $\delta = 0.9$.

3.1 Grim-trigger

This subsection examines a representative FSA that follows a well-known equilibrium strategy under perfect monitoring, called grim-trigger (GT), in which a player first cooperates, but as soon as she observes a defection, she defects forever. As shown in Fig. 1, this FSA has two states, i.e., R (reward) and P (punishment). Player i takes $a_i = C$ in state R and $a_i = D$ in state P. When both players act according to the GT FSA, a joint FSA has four states, i.e., RR, RP, PR, and PP. The system of linear equations for this joint FSA is given as

$$\begin{pmatrix} V_{RR} \\ V_{RP} \\ V_{PR} \\ V_{PP} \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \\ 2 \\ 0 \end{pmatrix} + \delta \begin{pmatrix} p & q & q & r \\ 0 & q+r & 0 & p+q \\ 0 & 0 & q+r & p+q \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} V_{RR} \\ V_{RP} \\ V_{PR} \\ V_{PP} \end{pmatrix}.$$

By solving this, we obtain

$$V_{RR} = \frac{1 - \delta\, r}{(1 - \delta\, p)\ (1 - \delta\, r - \delta\, q)}.$$

Figure 9 illustrates the range of signal parameters over which GT constitutes an SPFSE. The x-axis indicates p, the probability that signals are correct, e.g., $o((g, g)|(c, c))$ or $o((b, g)|(c, d))$. The y-axis indicates q, the probability that signals have exactly one error, e.g., $o((g, b)|(c, c))$ or $o((b, g)|(d, d))$. When p is large, the signals of the two players tend to be correct, e.g., the player is likely to observe g/b when her opponent cooperates/defects. When q is small, the signals are strongly correlated, i.e., if the signal of a player is wrong, the signal of her opponent is also likely to be wrong.

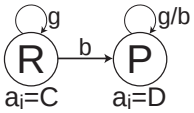

Figure 1: GT

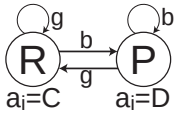

Figure 2: TFT

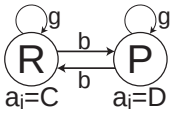

Figure 3: 1-MP

Basically, GT constitutes an SPFSE where p is large and q is small, i.e., the signals tend to be correct and strongly correlated. In the right-hand side of this region, p is large but q becomes larger, i.e., the signals tend to be correct but not strongly correlated. Then, let us assume player 1 observes b in a period. Player 1 is quite sure that this is an error. Furthermore, since the correlation is not so strong, player 2 is likely to receive a correct signal. Thus, for player 1, it is better to keep cooperation rather than to start defection. Also, in the left-hand side of this region, p is relatively small. Then, even after two periods, the probability that player 2 has been receiving correct signals in a row is not so large. Thus, for player 1, it is better to start defection rather than to keep cooperation.

A shortcoming of GT is that it is too unforgiving and thus generates a low payoff. For example, when $p = 0.9$, $q = 0.01$, and $\delta = 0.9$, the expected discounted payoff is about 5.31, while if players can keep cooperating, the expected discounted payoff would be 10. In the next section, we explore a class of strategies that is more forgiving.

4 k-period mutual punishment

A well-known, more forgiving strategy is tit-for-tat (TFT) (Fig. 2). However, if two players use TFT, an observation of defection leads to poorly coordinated behavior. Figure 4 shows the joint FSA for TFT. Thick/thin/dotted lines represent the transition with probabilities p, r, and q, respectively. Notice that we assume p is much larger than q or r. We can see that after an observation error players largely alternate between (C, D) and (D, C). In such a situation, a player is better off deviating to end this cycle and returns to (C, C). For this reason, TFT does not constitute an SPFSE. Note that, basically for the same reason, TFT does not constitute a subgame perfect equilibrium under perfect monitoring. Furthermore, the payoff associated with TFT is low. After an observation error, it is hard to go back to (C, C), as Figure 4 shows. In fact, the probability of (C, C) in the invariant distribution is 0.25, as long as $q > 0$ and $r > 0$.

Now, let us consider a slightly modified FSA in Fig. 3, which we call 1-period

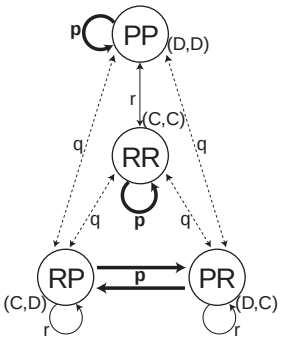

Figure 4: Joint FSA for TFT

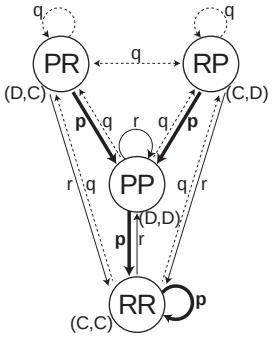

Figure 5: Joint FSA for 1-MP

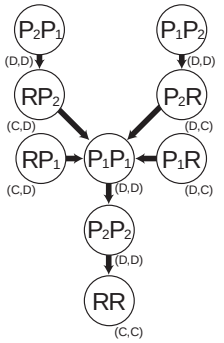

Figure 6: Joint FSA for 2-MP

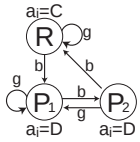

Figure 7: 2-MP

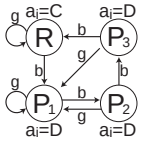

Figure 8: 3-MP

mutual punishment (1-MP). Traditionally, this FSA is known as *Pavlov* [Kraines and Kraines, 1989]. According to this FSA, a player first cooperates. If her opponent defects, she also defects, but after one period of mutual defection, she returns to cooperation. Figure 5 shows the joint FSA of 1-MP. We can see that after one observation error occurs players can quickly return to the mutual cooperation state RR. The expected probability (in the invariant distribution) that players are in state RR is about $p - 2q$.

Unfortunately, 1-MP does not constitute an SPFSE in our parameterization, since it is too forgiving. Basically, 1-MP punishes a deviator by one period of mutual defection. The gain from defection x is exactly equal to the loss in the next period y ($x = y = 1$). Therefore, as long as a player discounts future payoff, 1-MP cannot be an SPFSE, even under perfect monitoring. Consequently, we generalize the idea of 1-MP to k-period mutual punishment (k-MP). Under this FSA, a player first cooperates. If her opponent defects, she also defects, but after k consecutive periods of mutual defection, she returns to cooperation.

Figure 7 shows the FSA of 2-MP. 2-MP is less forgiving than 1-MP, since it cooperates approximately once in every three periods to the opponent who always defects. By increasing k, we can make this strategy less forgiving. When $k = \infty$, this strategy becomes equivalent to GT. Figure 6 shows a joint FSA for 2-MP. For simplicity, we only show thick lines that represent the transition with probability p. We can see that after some observation errors occur players can quickly return to the mutual cooperation state RR.

Figure 9 illustrates the range of signal parameters over which 2-MP is an SPFSE. For comparison, we show the range where GT is an SPFSE. We can see

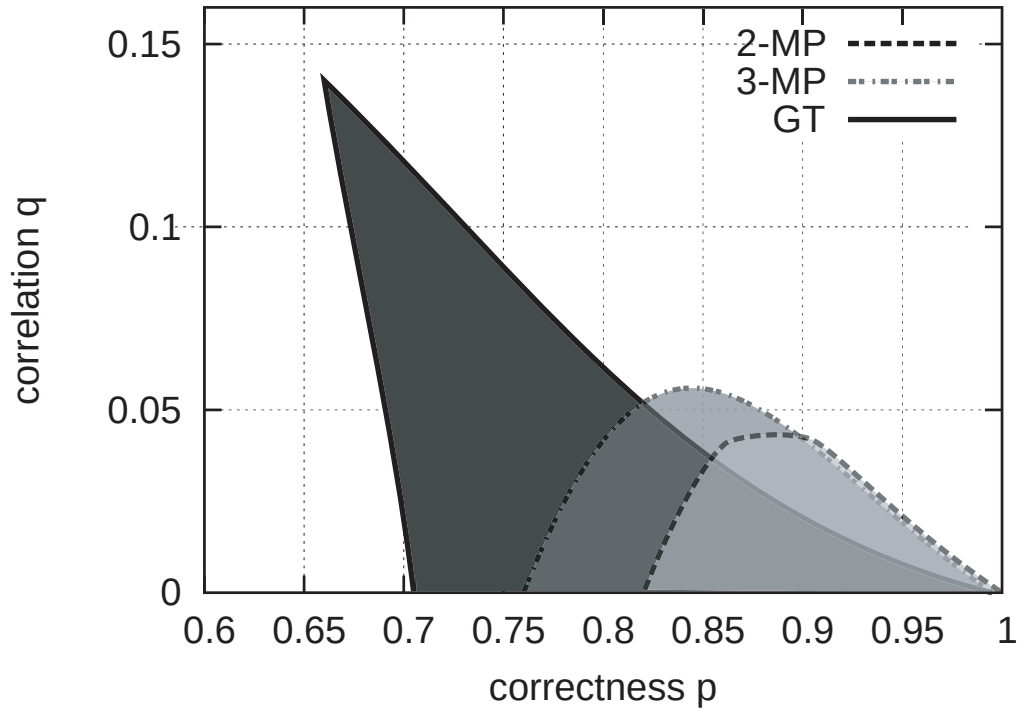

Figure 9: Range of signal parameters over which GT/2-MP/3-MP is an SPFSE. Note that feasible parameter space is $p + 2q \leq 1$.

that even for $k = 2$, k-MP can be an SPFSE in a reasonably wide range of signal parameters, though the size of the range is smaller than GT. When the correlation of signals is quite strong ($q \fallingdotseq 0$), 2-MP constitutes an SPFSE in the range of signal correctness $p \in [0.82, 1)$. As the correlation becomes slightly larger (i.e., $q > 0.04$), 2-MP is no longer an SPFSE. When $q = 0.04$, 2-MP constitutes an SPFSE in the range of correctness $p \in [0.86, 0.91)$. It is significant that GT is more sensitive to the correlation than 2-MP when p is sufficiently large. When the correctness p exceeds 0.86, there is a range of correlation where GT is not an SPFSE, but 2-MP is. Figure 9 also shows the range of signal parameters over which 3-MP is an SPFSE. The SPFSE range of 3-MP almost includes that of 2-MP.

Now, let us examine the average payoff of GT and k-MP. In Fig. 10, the x-axis indicates the correctness of signal p, while the correlation q is fixed at 0.01. The y-axis indicates the average payoff per period. Note that average payoff is 1 if mutual cooperation is always achieved. It is clear that 2-MP significantly

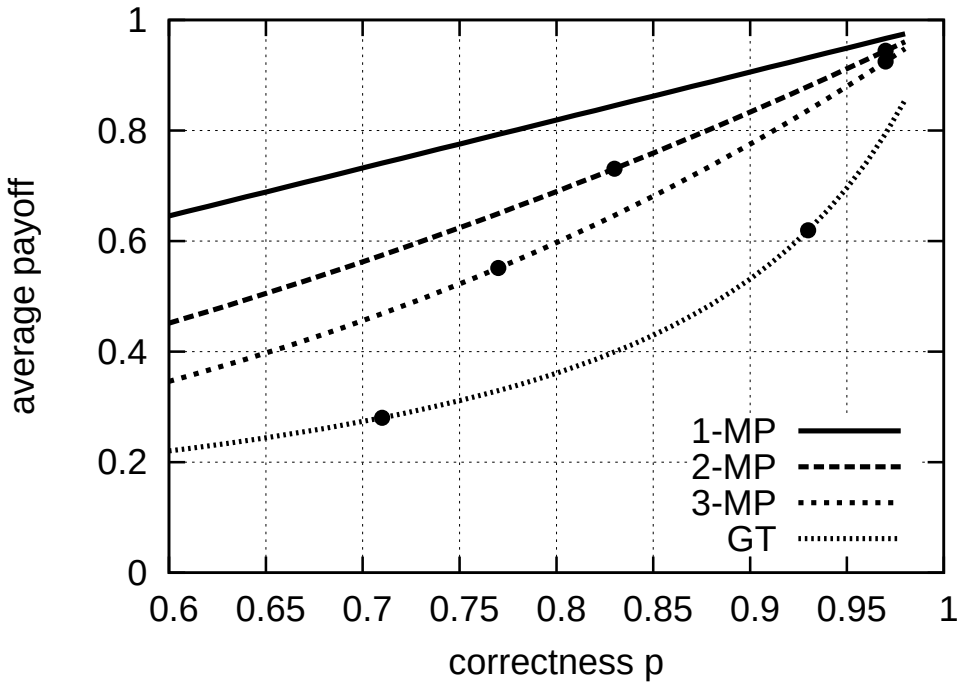

Figure 10: Average payoff per period of FSA (q=0.01).

outperforms GT regardless of signal correctness. Moreover, 2-MP outperforms 3-MP regardless of signal correctness. 1-MP always outperforms 2-MP as well, but 1-MP cannot constitute an SPFSE in our payoff matrix. We also show two points on each line. Within the range between two points, an FSA constitutes an SPFSE. We can see that the size of the range becomes wider by increasing k, but the efficiency becomes lower.

Let us examine the effect of discount factor δ. Figure 11 shows the average payoffs by varying discount factor $\delta \in [0.5, 0.95]$. We show one point on each line; when δ is larger than that point, an FSA constitutes an SPFSE. We can see that when δ increases, the average payoffs decrease, and the difference between k-MP and GT becomes more significant.

One obvious question is whether there is any FSA (except k-MP) that constitutes an SPFSE and achieves a better efficiency. To answer this question, we exhaustively search for small-sized FSAs that can constitute an equilibrium. We enumerate all possible FSAs with fewer than three states, i.e., $|A|^{|\Theta|} \cdot |\Theta|^{|\Theta| \cdot |\Omega|}$=5832 FSAs, and check whether they constitute an SPFSE. We found that only eleven FSAs (after removing equivalent ones) could be an SPFSE in a reasonably wide

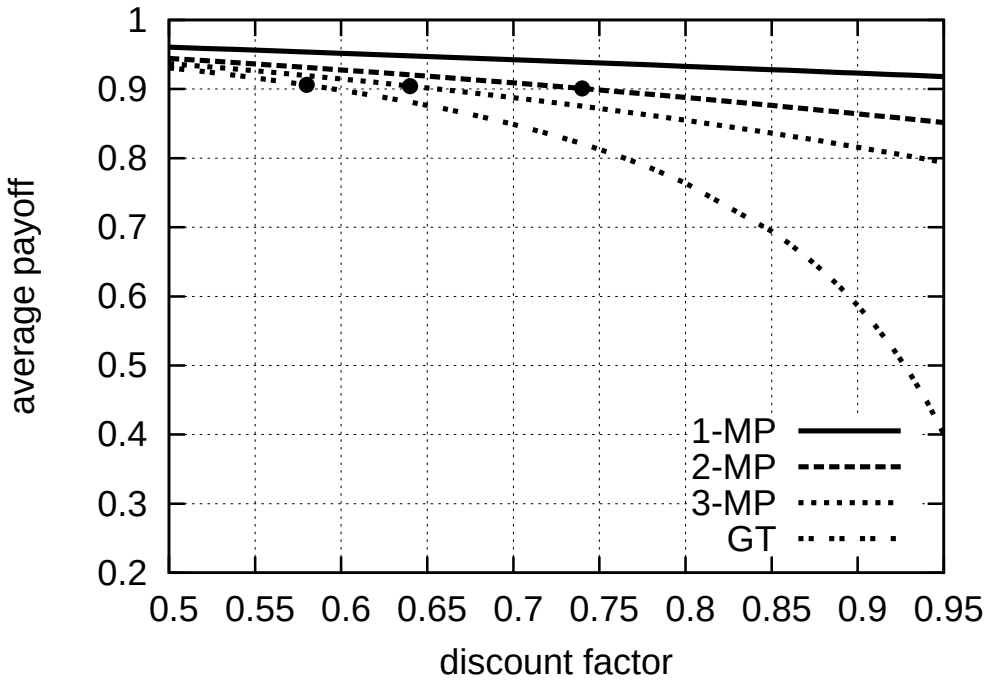

Figure 11: Effect of discount factors (p=0.92, q=0.01).

range of signal parameters. Furthermore, among them, 2-MP is the only FSA that is more efficient than GT.

5 Discussion

Traditionally, 1-MP has been called Pavlov or "win-stay, lose-shift [Nowak and Sigmund, 1993]." Pavlov is frequently used in the literature of evolutionary simulation, e.g., [Nowak, 2006]. They examine several extensions of Pavlov in repeated prisoner's dilemma where a player's action is subject to noise (*trembling hands*). It is well-known that Pavlov can constitute a subgame perfect Nash equilibrium in repeated prisoner's dilemma with perfect monitoring for a high discount factor. However, this is not well-investigated in the setting of private monitoring. To the best of our knowledge, 1-MP/Pavlov has not yet been identified as an equilibrium in repeated games with private monitoring.

Our equilibrium analysis suggests that 1-MP can be a more efficient equilibrium than 2-MP if the gain from deviation and the observation error are small

enough. In our payoff matrix, the gain from deviation in a period is as large as the loss from deviation in the next period. Thus, 1-MP cannot be an SPFSE. In our ongoing project, we have confirmed that 1-MP can constitute an SPFSE when the gain from deviation x is lower than or equal to 0.8 in some parameter settings of private signals.

We assumed a monitoring structure that can be naturally interpreted as a noisy signal about the opponent's action. In our setting, after (C, D), for example, players are most likely to observe (b, g). In contrast, the few existing works that found simple pure strategy equilibria under private monitoring [Mailath and Morris, 2002; Phelan and Skrzypacz, 2009] mostly concern the *almost public* monitoring case, where players are always likely to get the *same* signal (after (C, D), for example, players are likely to get (g, g) or (b, b)). Our preliminary study shows that 2-MP is no longer an SPFSE in such a case. We are planning to analyze this case using our POMDP approach.

6 Conclusions

This paper identifies simple pure strategy equilibria in an infinitely repeated prisoner's dilemma game with a noisy signal about the opponent's action, which is one instance of private monitoring. This analysis has been considered to be extremely hard. However, by utilizing a POMDP solver, we can check whether an FSA, which is a compact representation of an equilibrium behavior, constitutes an SPFSE. We first examine how observation errors affect the behavior of GT. Then we propose the k-MP strategy, which incorporates GT and Pavlov as a special case and show k-MP constitutes an SPFSE in a reasonably wide range of observation errors. Its efficiency is better than GT. We exhaustively search for simple FSAs with at most three states and confirm that no other FSA constitutes an equilibrium in a reasonably wide range of signal parameters and is more efficient than GT.

Our technique established in this paper suggests a host of new research agenda in repeated games with private monitoring. For example, we would like to investigate almost public monitoring cases where players are likely to observe a common signal regardless of their actions. Our preliminary investigation reveals that 2-MP does not constitute an SPFSE in such settings. In addition, we hope to investigate other games such as congestion games, which can model various application problems including a packet routing problem.

References

[Doshi and Gmytrasiewicz, 2006] Doshi, P., and Gmytrasiewicz, P. J. 2006. On the Difficulty of Achieving Equilibrium in Interactive POMDPs. In *AAAI*.

[Kaelbling, Littman, and Cassandra, 1998] Kaelbling, L. P.; Littman, M. L.; and Cassandra, A. R. 1998. Planning and acting in partially observable stochastic domains. *Artificial Intelligence* 101:99–134.

[Kandori and Obara, 2010] Kandori, M., and Obara, I. 2010. Towards a Belief-Based Theory of Repeated Games with Private Monitoring: An Application of POMDP. mimeo.

[Kandori, 2010] Kandori, M. 2010. Repeated games. *Game theory*. Palgrave macmillan. 286–299.

[Kraines and Kraines, 1989] Kraines, D., and Kraines, V. 1989. Pavlov and the prisoner's dilemma. *Theory and Decision* 26:47–79.

[Mailath and Morris, 2002] Mailath, G. J., and Morris, S. 2002. Repeated games with almost-public monitoring. *Journal of Economic Theory* 102(1):189–228.

[Nowak and Sigmund, 1993] 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.

[Nowak, 2006] Nowak, M. 2006. *Evolutionary Dynamics: Exploring the Equations of Life*. Harvard University Press.

[Phelan and Skrzypacz, 2009] Phelan, C., and Skrzypacz, A. 2009. Beliefs and Private Monitoring. mimeo.

[Russell and Norvig, 2002] Russell, S., and Norvig, P. 2002. *Artificial Intelligence: A Modern Approach (2nd Edition)*. Prentice Hall.

[Tennenholtz and Zohar, 2009] Tennenholtz, M., and Zohar, A. 2009. Learning equilibria in repeated congestion games. In *AAMAS*, 233–240.

[Wang, Chatterjee, and Kwiat, 2009] Wang, W.; Chatterjee, M.; and Kwiat, K. 2009. Cooperation in ad hoc networks with noisy channels. In *6th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks*, 1–9.