Automated Equilibrium Analysis of Repeated Games with Private Monitoring: A POMDP Approach

Atsushi Iwasaki, Kyushu University
Yongjoon Joe, Kyushu University
Michihiro Kandori, University of Tokyo
Ichiro Obara, UCLA
Makoto Yokoo, Kyushu University

The present paper investigates repeated games with *imperfect private monitoring*, where each player privately receives a noisy observation (signal) of the opponent's action. Such games have been paid considerable attention in the AI and economics literature. Since players do not share common information in such a game, characterizing players' optimal behavior is substantially complex. As a result, identifying pure strategy equilibria in this class has been known as a hard open problem. In our previous work, we showed that the theory of partially observable Markov decision processes (POMDP) can be applied to identify a class of equilibria where the equilibrium behavior can be described by a finite state automaton (FSA). However, we have not yet provide a practical method or a program to apply this general idea to actual problems. In this paper, we first develop a program that acts as a wrapper of a standard POMDP solver, which takes a description of a repeated game with private monitoring and an FSA as inputs, and automatically checks whether the FSA constitutes a symmetric equilibrium. We apply our program to repeated Prisoner's dilemma and find a novel class of FSA, which we call k-period mutual punishment (k-MP). The k-MP starts with cooperation and defects after observing a defection. It restores cooperation after observing defections k-times in a row. Our program enables us to exhaustively search for all FSAs with at most three states, and we found that 2-MP beats all the other pure strategy equilibria with at most three states for some range of parameter values and it is more efficient in an equilibrium than the grim-trigger.

Categories and Subject Descriptors: I.2.11 [Artificial Intelligence]: Distributed Artificial Intelligence—*Multi-agent systems*; J.4 [Social and Behavioral Sciences]: Economics

General Terms: Algorithms, Economics, Theory

Additional Key Words and Phrases: Game theory, repeated games, private monitoring, POMDP

ACM Reference Format:
Iwasaki, A., Joe, Y., Kandori, J., Obara, I., Yokoo, M. 2012. Automated Equilibrium Analysis of Repeated Games with Private Monitoring: A POMDP Approach ACM Trans. Embedd. Comput. Syst. 9, 4, Article 39 (March 2010), 17 pages.
DOI = 10.1145/0000000.0000000 http://doi.acm.org/10.1145/0000000.0000000

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 relationships among players and has a wide range of applications, e.g., secret price cutting and agent planning under uncertainty. Therefore,

Author's addresses: A. Iwasaki, Y. Joe, M. Yokoo, Graduate School of ISEE, Kyushu University; M. Kandori, Faculty of Economics, University of Tokyo; I. Obara, Department of Economics, UCLA.

DOI 10.1145/0000000.0000000 http://doi.acm.org/10.1145/0000000.0000000

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, Ng and Seah [2008] examine protocols in multihop wireless networks with self-interested agents and Wang et al. [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.

Analytical studies on this class of games have not been quite successful. The difficulty comes from the fact that players do not share common information under private monitoring, and finding pure strategy equilibria in such games has been known as a hard open problem [Mailath and Samuelson 2006]. Under private monitoring, each player cannot observe the opponents' private signals, and he or she has to draw statistical inferences about the history of the opponents' private signals. The inferences quickly become very complicated over time, even if players adopt relatively simple strategies [Kandori 2010]. As a result, finding a profile of strategies which are mutual best replies after any history, i.e., finding an equilibrium, is a quite demanding task.

In our previous work [Kandori and Obara 2010], we showed that the theory of the partially observable Markov decision process (POMDP) can be used to identify equilibria, when equilibrium behavior can be described by a finite state automaton (FSA). We believe this result is significant since it implies that by utilizing a POMDP solver, we can systematically determine whether a given profile of FSAs can constitute an equilibrium. Furthermore, this result is interesting since it connects two popular areas in AI and multi-agent systems, namely, POMDP and game theory.

Traditionally, in the AI literature, the POMDP framework is a popular approach for single-agent planning/control, and game theory has been extensively used for analyzing multi-agent interactions. However, these two areas have not been well-connected so far, as mentioned in the most recent edition of a popular AI textbook "... game theory has been used primarily to analyze environments that are at equilibrium, rather than to control agents within an environment" [Russell and Norvig 2009]. As one notable exception, Doshi and Gmytrasiewicz [2006] investigate the computational complexity and subjective equilibrium. In a subjective equilibrium, a player may not perfectly know the opponent's strategy. As a result, the definition of a subjective equilibrium is involved, and they showed that reaching a subjective equilibrium is difficult under the limit of computational complexity. In contrast, we have examined that if simple behavior described by FSA can be mutual best replies and proposed a general method to check if a given profile of FSAs constitutes an equilibrium [Kandori and Obara 2010].

Also, Hansen et al. [2004] deal with partially observable stochastic games (POSGs) and develop an algorithm that iteratively eliminates dominated strategies POSGs can be considered a generalization of repeated games with private monitoring, since agents might play different games at each stage. However, this algorithm can be applied only for a finite horizon, and it cannot guarantee to identify an equilibrium.

Furthermore, in our previous work [Nair et al. 2003], we present a decentralized POMDP algorithm called Joint Equilibrium-based Search for Policy with Nash Equilibrium (JESP-NE). In this algorithm, a locally optimal joint policy for cooperative agents is obtained by utilizing a POMDP solver. However, in this approach, we restrict our attention to finite horizon cases and assume a policy/strategy is not necessarily represented as an FSA. Also, in our another previous work [Marecki et al. 2008], we consider infinite horizon cases and assume a policy/strategy is represented as an FSA. However, in this approach, the goal is to find a joint FSA that obtains the best reward as a team. Thus, we do not guarantee that the joint FSA constitutes an equilibrium.

Unfortunately, our previous results [Kandori and Obara 2010] have not yet been widely acknowledged in the AI and agent research communities. Furthermore, for the time being, there exists no work that actually applies this method to identify equilibria of repeated games even in the economics/game-theory field.

The main difficulty for utilizing the result is that, although a general theoretical idea is presented based on POMDP, to identify equilibria of repeated games with private monitoring, we have not demonstrated how to implement our idea computationally. Moreover, it has not yet been confirmed that this approach is really feasible when analyzing problem instances that are complex enough to represent realistic and meaningful application domains. In particular, we found that there exist one non-trivial difference between the POMDP model and the model for repeated games with private monitoring. More precisely, in a standard POMDP model, we usually assume that an observation depends on the current action and the next state. On the other hand, in the model of repeated games, we assume that an observation depends on the current action and the current state. As a result, applying/extending the results of Kandori and Obara [2010] is difficult for researchers in game theory, as well as those in the AI and agent research communities.

To overcome this difficulty, we first develop a program that acts as a wrapper of a standard POMDP solver. This program takes a description of a repeated game with private monitoring and an FSA as inputs. Then, this program automatically creates an input for a POMDP solver, by taking into account the differences in the models described above. Next, this program runs a POMDP solver, analyzes the obtained results, and answers whether the FSA constitutes a symmetric equilibrium.

Furthermore, as a case study to confirm the usability of this program, we identify equilibria in an infinitely repeated prisoner's dilemma game, where each player privately receives a noisy signal about each other's actions. First, we consider the situation where an opponent's action is observed with small observation errors. This case is referred to as the *nearly-perfect* monitoring case. Although the monitoring structure is quite natural, systematically finding equilibria in such structure has not been possible without utilizing a POMDP solver. We exhaustively search for simple FSAs with a small number of states and find a novel class of FSA called k-period mutual punishment (k-MP). Under this FSA, a player first cooperates. If she observes a defection, she also defects, but after the observation of k consecutive defections, she returns to cooperation. We can control the forgiveness of k-MP by changing the parameter k. Note that k-MP incorporates grim-trigger and the well-known strategy *Pavlov* [Kraines and Kraines 1989] as a special case ($k = \infty$ or $k = 1$). Although it is somewhat counterintuitive, requiring such mutual defection periods is beneficial in establishing a robust coordination among players in the nearly-perfect monitoring case. In contrast, in the *almost-public* monitoring case, the tit-for-tat (TFT) can better coordinate players' behavior; TFT can be an equilibrium, while k-MP is not. In both cases, the grim-trigger can be an equilibrium. Accordingly, our program helps us to gain important insights into the way players coordinate their behavior under different private monitoring structures.

2. REPEATED GAMES WITH PRIVATE MONITORING

2.1. Model

We model a repeated game with private monitoring. 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. This model is the standard one in the repeated game literature with private monitoring [Kandori and Obara 2010].

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 $(1 - \delta)G_i$.

2.2. Repeated game strategies and finite state automata

We now explore several ways to represent repeated game strategies. We start with the conventional representation of strategies in the repeated game defined above. 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, $s_i : H_i \rightarrow A$, 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, $f : \Theta \rightarrow A$ determines the action choice for each state, and $T : \Theta \times \Omega \rightarrow \Theta$ 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., $m = \langle \Theta, f, T \rangle$, a finite state *preautomaton* (pre-FSA). Now, we introduce a *symmetric pure finite state equilibrium*.

Definition 2.1. 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 a Nash equilibrium for dynamic games of imperfect information. Traditionally, this concept is defined for a *full* repeated game strategy, i.e., a strategy must specify actions for all possible histories including histories for off-equilibrium paths. As a result, an strategy tends to be quite complex; we might need an infinite number of states to represent such a strategy using an FSA. Thus, analytical studies on this class of games have not been quite successful so far. In contrast, our definition requires that an FSA specifies only the behavior of a player on equilibrium paths. As a result, we can concisely represent an equilibrium strategy in our definition.

It must be emphasized that if an FSA M constitutes an equilibrium, it means that as long as player 2 acts 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 (please consult Example 1 in [Kandori and Obara 2010] for details).

2.3. Monitoring structures in repeated prisoner's dilemma

We apply the POMDP technique to the prisoner's dilemma model. The stage game payoff is given as follows.

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

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 $\omega_i = b$. Let us introduce the joint distribution of private signals $o(\boldsymbol{\omega} \mid \boldsymbol{a})$ for the prisoner's dilemma model. When the action profile is (C, C), the joint distribution is given as follows (when the action profile is (D, D), p and s are exchanged).

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

Notice that the probability that players 1 and 2 observe (g, g) is p, and the probability that they observe (g, b) is q.

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), v and u are exchanged).

	$w_2 = g$	$w_2 = b$
$w_1 = g$	t	u
$w_1 = b$	v	w

These joint distributions of private signals require only the constraints of $p+q+r+s = 1$ and $t + u + v + w = 1$.

Repeated games with private monitoring is a generalization of infinitely repeated games with conventional imperfect monitoring. By changing signal parameters, the joint distributions can represent any monitoring structure in repeated games. Let us briefly explain several existing monitoring structures. First, we say monitoring is *perfect* if each player perfectly observes the opponent's action in each period, i.e., $p = v = 1$ and $q = r = s = t = u = w = 0$ hold. Second, we say monitoring is *public* if each player always observes a common signal, i.e., $p + s = t + w = 1$ and $q = r = u = v = 0$ hold. Third, we say monitoring is *almost-public* if players are always likely to get the *same* signal (after (C, D), for example, players are likely to get (g, g) or (b, b)), i.e., $p + s = t + w \approx 1$ and $q = r = u = v \approx 0$.

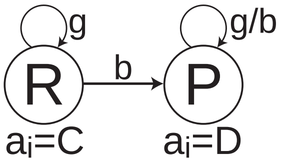

Fig. 1. GT

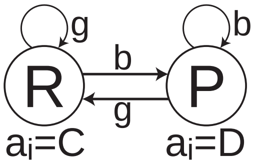

Fig. 2. TFT

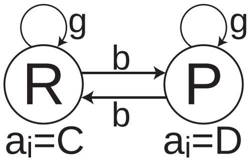

Fig. 3. 1-MP

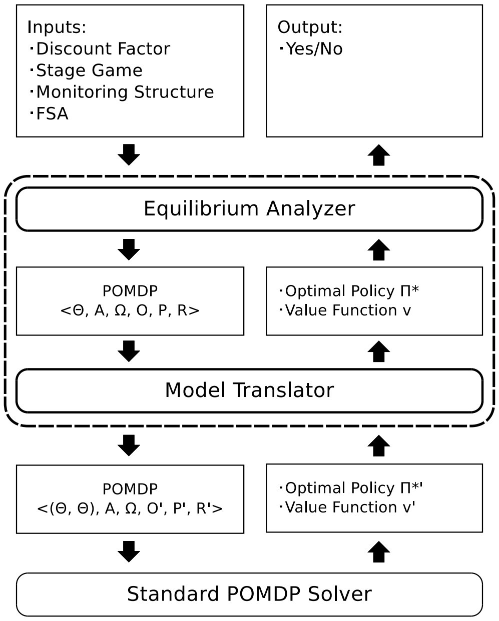

Fig. 4. Equilibrium analyzer

2.4. Existing FSAs

Let us summarize the existing FSAs in the literature of repeated games. First, grim-trigger (GT) is a well-known FSA under which a player first cooperates, but as soon as she observes 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. GT can often constitute an equilibrium under perfect and imperfect monitoring.

Second, tit-for-tat (TFT) is another well-known FSA in Fig. 2. It is well known that TFT does not prescribe mutual best replies after a deviation (hence it is *not* a subgame perfect Nash equilibrium) under perfect monitoring. This problem does not arise under public and almost-public monitoring, and TFT can be a sequential equilibrium under public monitoring.

Finally, 1-period mutual punishment (1-MP) in Fig. 3 is known as *Pavlov* [Kraines and Kraines 1989], *simpleton* [Rapoport and Chammah 1965], *perfect tit-for-tat* [Fudenberg and Tirole 1991], or *win-stay, lose-shift* [Nowak and Sigmund 1993]." 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.

Pavlov is frequently used in the literature of evolutionary simulation, e.g., [Kraines and Kraines 1989; Nowak and Sigmund 1993]. They examine several extensions of Pavlov in the 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 under perfect monitoring. However, this has not been investigated well 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. We will again discuss TFT and 1-MP under our monitoring structures in Section 4.

3. PROGRAM FOR EQUILIBRIUM ANALYSIS

In this section, we describe our newly developed program that checks whether an FSA $M = \langle \Theta, \hat{\theta}, f, T \rangle$ constitutes an SPFSE according to Fig. 4.[1]

3.1. Main Procedure

Let us describe the main procedures of our program indicated as "Equilibrium Analyzer" and "Standard POMDP solver" in Fig. 4. 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.

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

Now, let us consider how to obtain the best response for player 1, assuming player 2 acts 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: $O(\omega_1 \mid a_1, \theta^t) = o_1(\omega_1 \mid (a_1, f(\theta^t)))$.

Note that in a standard POMDP model, we usually assume that the observation probability depends on the next state θ^{t+1} rather than on the current state θ^t. We present this alternative model here, since it is more suitable for representing repeated games with private monitoring. In the next subsection, we show how to map this model into the standard formulation of POMDP.

$P(\theta^{t+1} \mid \theta^t, a_1)$ 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:

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

An expected payoff function $R : A \times S \to \mathbb{R}$ 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 general ideas presented in our previous work [Kandori and Obara 2010], but this description is 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
$\langle \Theta, A, \Omega, O, P, R \rangle$. Since our POMDP model is different from the standard POMDP model, we cannot directly use a standard POMDP solver such as [Kaelbling et al. 1998]. We describe how to absorb this difference in the next subsection. In general,

[1] Our software will be publicly available.

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.[2]

(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 FSA 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. To overcome this problem, we use m as an initial policy and make sure that Π^* includes m as long as M constitutes an SPFSE.

3.2. Procedure for Handling Model Differences

In this subsection, corresponding with “Model Translator” in Fig. 4, we describe a method for translating a POMDP description $\langle \Theta, A, \Omega, O, P, R \rangle$ in our model, into a standard model $\langle \Theta', A, \Omega, O', P', R' \rangle$. Here, the possible set of actions A and observations Ω are the same in these two models.

The key idea of this translation is to introduce a set of new combined states Θ', where $\Theta' = \Theta^2$. Namely, we assume that a state θ'^t in the standard POMDP model represents the combination of the previous and current states (θ^{t-1}, θ^t) in our model present in the previous subsection. For example, assume player 1 acts according to an FSA called grim-trigger (GT) defined in Fig. 1. There are two states in the original model. Consequently, in the standard model, there are $2 \times 2 = 4$ states, i.e., $\Theta' = \{(R,R),(R,P),(P,R),(P,P)\}$. Among these four states, (P,R) is infeasible, and thus there exists no state transition to (P,R).

A new state transition function $P'(\theta'^{t+1} \mid \theta'^t, a_1)$ is equal to $P(\theta^{t+1} \mid \theta^t, a_1)$ in the original model if $\theta'^{t+1} = (\theta^t, \theta^{t+1})$ and $\theta'^t = (\theta^{t-1}, \theta^t)$, i.e., the previous state in θ'^{t+1} and the current state in θ'^t are identical. Otherwise, it is 0. Next, let us examine how to define $O'(\omega_1 \mid a_1, (\theta^t, \theta^{t+1}))$. This is identical to the posterior probability that the observation was ω_1, when the state transits from θ^t to θ^{t+1}. Thus, this is defined as:

$$O'(\omega_1 \mid a_1, (\theta^t, \theta^{t+1})) = \frac{\sum_{\omega_2 \in \Omega'} O(\omega_1, \omega_2 \mid (a_1, f(\theta^t)))}{\sum_{\omega \in \Omega} \sum_{\omega_2 \in \Omega'} O(\omega, \omega_2 \mid (a_1, f(\theta^t)))},$$

where $\Omega' = \{\omega_2 \mid T(\theta^t, \omega_2) = \theta^{t+1}\}$. For example, let us consider that player 1 takes $a_1 = C$ when player 2, who acts according to GT, is in state (R,R). The probability that player 1 observes $w_1 = g$ is given by

$$O'(g \mid C, (R,R)) = \frac{O(g,g \mid (C,C))}{O(g,g \mid (C,C)) + O(b,g \mid (C,C))}.$$

[2] When the obtained policy is semi-optimal but $v(b_{\hat{\theta}}) = V_{\hat{\theta},\hat{\theta}}$ holds, we run a procedure to check $v(b_{\hat{\theta}})$ remains the same in an optimal, non-FSA policy.

Finally, the expected payoff function, $R'(a_1, (\theta^{t-1}, \theta^t))$, is given as $R(a_1, \theta^t)$.

This translation does not affect the optimal policy. More specifically, by solving the translated POMDP
$\langle \Theta', A, \Omega, O', P', R' \rangle$, we obtain an optimal policy Π'^* (which is described as a pre-FSA) and its value function $v'(b_{\theta'})$. Then, an optimal policy Π^* of the original POMDP
$\langle \Theta, A, \Omega, O, P, R \rangle$ is identical to Π'^*. Also, from $b_{\theta'}$, which is a belief over $\theta' = (\theta^{t-1}, \theta^t)$, we can extract b_{θ^t}, i.e., a belief over the current state. Then, $v'(b_{\theta'}) = v(b_{\theta^t})$ holds.

3.3. Program Interface

This program takes the discount factor, the description of a stage game, a monitoring structure defined by $o(\boldsymbol{\omega} \mid \boldsymbol{a})$, i.e., the probability of private signal profile $\boldsymbol{\omega}$ given an action profile $\boldsymbol{a}$, and an FSA, as “Inputs” of Fig. 4. Let us show an example. The meanings of these descriptions are self-explanatory.

```
discount: 0.9
actions: C D
# payoff matrix
PM:C:C: 1: 1
PM:D:C: 2:-1
PM:C:D:-1: 2
PM:D:D: 0: 0

observations: g b
# observation probability
O:g:g:C:C:0.97
O:b:g:C:C:0.01
O:g:b:C:C:0.01
O:b:b:C:C:0.01
...
# FSA description of Grim-trigger
states: R P
start: R
T:R:g:R
T:R:b:P
T:P:g:P
T:P:b:P
```

4. REPEATED PRISONER’S DILEMMA WITH NOISY OBSERVATION

This section first defines a monitoring structure that is *nearly-perfect*. We say monitoring is nearly-perfect if each player is always likely to perfectly observe the opponent’s action in each period, i.e., $p = v$, $q = r = t = w$, and $s = u = 1 - p - 2q$, where p is much larger than q or s. We assume $p \in (1/2, 1)$ and $q \in (0, 1/4)$ under the constraint $p + 2q + s = 1$. Although the monitoring structure is quite natural, systematically finding equilibria in such structure has not been possible without utilizing a POMDP solver.

In addition to nearly-perfect monitoring, we also consider almost-public monitoring. Recall that, under almost-public monitoring, players are always likely to get the same signal. Thus, we set our parameters as follows: $p + s = t + w \approx 1$ and $q = r = u = v \approx 0$. We assume $p \in (1/2, 1)$, $q \in (0, 1/4)$, and $t \in (0, 1/2)$ under the constraints $p + 2q + s = 1$ and $t + 2q + w = 1$.

Notice that $\pi_i(a_i, \omega_i)$ is chosen so that $g_i(\boldsymbol{a})$ is constant under both monitoring structures. Throughout our paper, we use the default setting: $x = 1$, $y = 1$, and the discount factor $\delta = 0.9$. Next, this section identifies signal parameters where GT, TFT, and 1-MP constitute an SPFSE according to our program.

4.1. Grim-trigger

This subsection examines a representative FSA, called grim-trigger (GT). When both players act according to GT, a joint FSA has four states: RR, RP, PR, and PP. Under nearly-perfect monitoring, the system of linear equations for this joint FSA is given as

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

By solving this, we obtain

$$V_{RR} = \frac{1-\delta\, s}{(1-\delta\, p)\ (1-\delta\, s-\delta\, q)}, V_{PR} = -\frac{2}{\delta\, s+\delta\, q-1}, V_{RP} = \frac{1}{\delta\, s+\delta\, q-1},\ \text{and } V_{PP} = 0.$$

Figure 12 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.

Basically, GT constitutes an SPFSE where p is large and q is small, i.e., the signals are accurate and strongly correlated. Suppose p is large but q is not small, and assume that player 1 observes b. 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 deviate from GT and to keep cooperation. When p is relatively small, in contrast, the probability that the opponent observes b is large. Therefore, it is better to start with defection. 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.

4.2. TFT and 1-MP

TFT in Fig. 2 is well-known as a more forgiving strategy than GT. However, if two players use TFT, an observation of defection leads to poorly coordinated behavior. Figure 5 shows the joint FSA for TFT under nearly-perfect monitoring. Thick/thin/dotted lines represent the transition with probabilities p, q, and s, respectively. Notice that we assume p is much larger than q or s. 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 returning to (C,C). For this reason, TFT does not constitute an SPFSE under nearly-perfect monitoring. Note that, basically for the same reason, TFT does not constitute a subgame perfect Nash equilibrium under perfect monitoring. Furthermore, the payoff associated with TFT is low. After an observation error, it is difficult to go back to (C,C), as Fig. 5 shows. In fact, the probability of (C,C) in the invariant distribution is 0.25, as long as $q > 0$ and $s > 0$.

Let us turn our attention to almost-public monitoring for a moment. We examined whether TFT is an SPFSE or not under almost-public monitoring within a wide range of signal parameters by utilizing our developed software. We confirmed TFT is an SPFSE only if q is smaller than about 0.07 in our parameterization. If two players

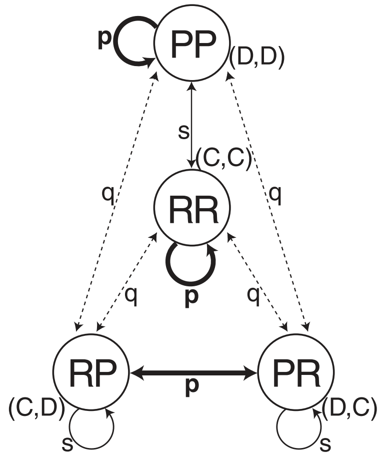

Fig. 5. Joint FSA for TFT under nearly-perfect monitoring

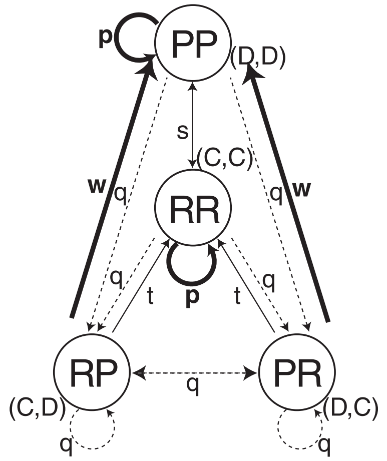

Fig. 6. Joint FSA for TFT under almost-public monitoring

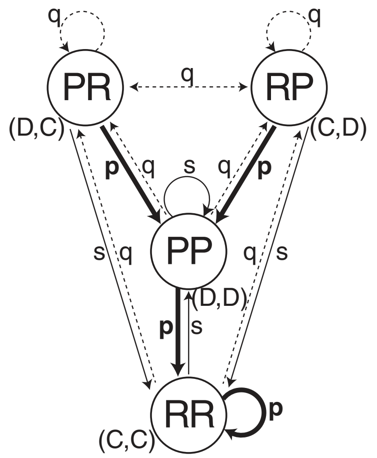

Fig. 7. Joint FSA for 1-MP under nearly-perfect monitoring

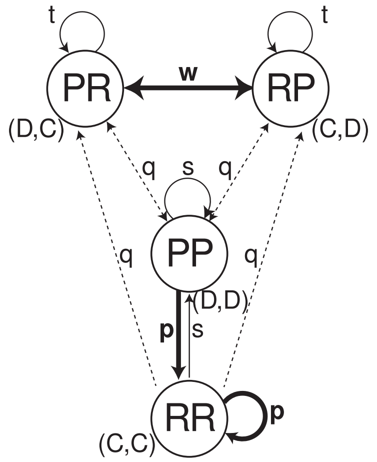

Fig. 8. Joint FSA for 1-MP under almost-public monitoring

use TFT under almost-public monitoring, an observation of defection leads to coordinated behavior. Figure 6 shows the joint FSA. Thick/thin/dotted lines represent the transition with probabilities $p\ (w)$, $s\ (t)$, and q, respectively. We can see that after an observation error players no longer alternate between (C, D) and (D, C). Although they will likely transit or stay at the mutual punishment state PP, they are more likely to return to the mutual cooperation state RR than under nearly-perfect monitoring. Notice that the similar argument can be applied to the public monitoring case.

The fact that TFT can be an SPFSE under almost-public monitoring has already been shown by Phelan and Skrzypacz [2012]. However, their analysis is limited to only very restricted parameter settings. Our software enables us to systematically search a variety of parameter settings. Also, we exhaustively search for all FSAs with at most three states that can constitute an equilibrium under almost-public monitoring, and found that TFT is the most efficient in SPFSE among these FSAs, including GT.

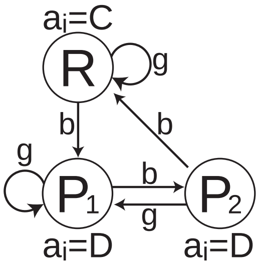

Fig. 9. 2-MP

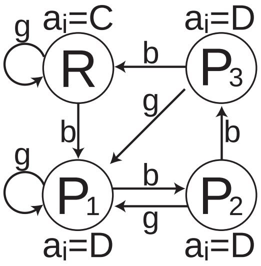

Fig. 10. 3-MP

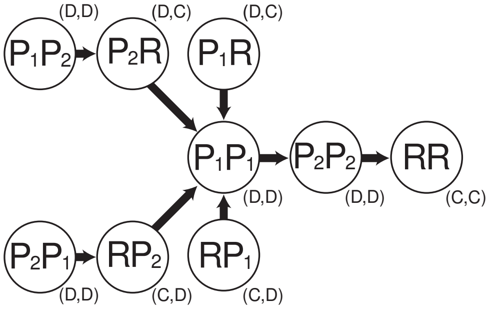

Fig. 11. Joint FSA for 2-MP under nearly-perfect monitoring

Now, let us consider the FSA in Fig. 3, which we call 1-period mutual punishment (1-MP). As we noted, traditionally, this FSA is known as Pavlov [Kraines and Kraines 1989]. Recall that, 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 7 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.[3] Also, 1-MP does not constitute an SPFSE under almost-public monitoring. Figure 8 illustrates that an observation of defection leads to poorly coordinated behavior, as in TFT under nearly-perfect monitoring.

5. K-PERIOD MUTUAL PUNISHMENT

This section generalizes 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 9 shows the FSAs 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 11 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

[3] 1-MP is a subgame perfect Nash equilibrium under perfect monitoring only if $\frac{1+x}{1-\delta^2} < \frac{1}{1-\delta}$.

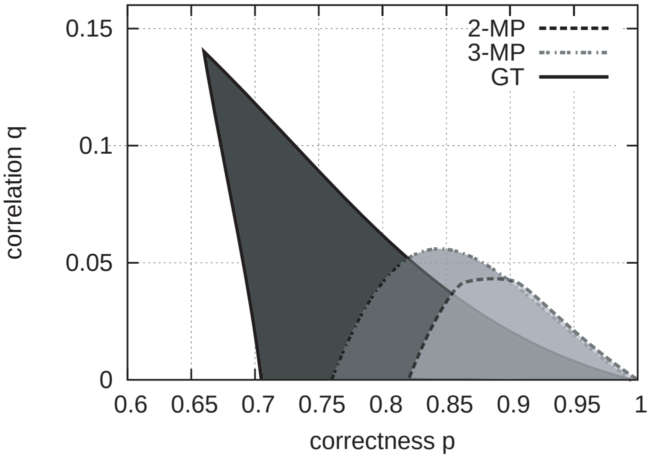

Fig. 12. Range of signal parameters over which GT/2-MP/3-MP is an SPFSE. Note that feasible parameter space is $p + 2q \leq 1$.

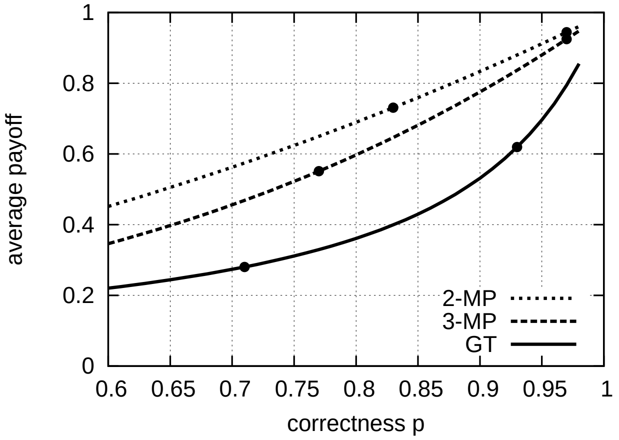

Fig. 13. Average payoff per period of FSA ($q = 0.01$).

some observation errors occur players can quickly return to the mutual cooperation state RR.

Figure 12 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 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 12 also shows the range of signal parameters over which 3-MP (Fig. 10) is an SPFSE. The SPFSE range of 3-MP includes almost all that of 2-MP.

Now, let us examine the average payoff of GT and k-MP. In Fig. 13, 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 outperforms GT and 3-MP regardless of signal correctness. We also placed two points on each line. Within the range between the 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.

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 at most 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 range of signal parameters.

Next, let us consider a generalization of TFT called *tit-for-k-tats* (TF-k-T), which is based on a similar idea to k-MP. According to TF-k-T, a player first cooperates. If her opponent defects, she also defects, but after k periods of her opponent's cooperation, she returns to cooperation. Note that TF-k-T incorporates GT and TFT as a special case. Phelan and Skrzypacz [2012] have shown that TF-2-T can constitute an equilibrium under almost-public monitoring. We confirmed that TF-2-T is an SPFSE only if q is very small (about 0.05 or smaller) in our parameterization. Furthermore, our exhaustive search found that TF-1-T, a.k.a. TFT is the most efficient among all FSAs within three states. We also observed that as k increases, the range of signal parameters over which TF-k-T is an SPFSE becomes wider, but the average payoff becomes lower. This trend is similar to k-MP under nearly-perfect monitoring. Furthermore, we confirmed that k-MP is no longer an SPFSE under almost-public monitoring. This fact can be explained using the same argument presented in Section 4.2.

There has been a series of iterated prisoners' dilemma competitions in noisy environments [Rogers et al. 2007]. In these competitions, a player is assumed to make an error, i.e., she sometimes takes an action that is different from her intended action. However, players can publicly observe their realized actions. Thus, this monitoring structure is substantially different from both of our settings, i.e., nearly perfect and almost public.

Alternatively, these competitions are not for finding strategies that constitute an equilibrium. They only examines which program/strategy performs better within a limited set of programs/strategies in a round-robin style tournament. In contrast, we identified several FSA, e.g., 2-MP, that constitutes an equilibrium. The implication that an FSA constitutes an equilibrium is far-reaching. If 2-MP constitutes an equilibrium, as long as the opponent is playing 2-MP, the best strategy is also to play 2-MP. Playing any other strategy, including very sophisticated strategies considered in these competitions, is meaningless.

6. EXTENSION WITH A RANDOM PRIVATE SIGNAL

Let us assume that agents can observe additional signals which (i) do not affect payoffs, (ii) convey no information about players' actions, and (iii) are strongly correlated. Interestingly, players can achieve better coordination by utilizing such "irrelevant" almost public signals. More specifically, let us assume that a player observes whether a particular event happens or not before each stage game. We assume with probability p', that both players observe the event, with probability s' that neither players observes the event, and with probability $(1 - p' - s')/2$ that player 1 or 2 observes the event but player 2 or 1 does not, respectively. We assume p' is relatively small (not too frequent), and $(1 - p' - s')/2$ is much smaller than p', i.e., if one player observes the event, it is very likely that the other player also observes the event.

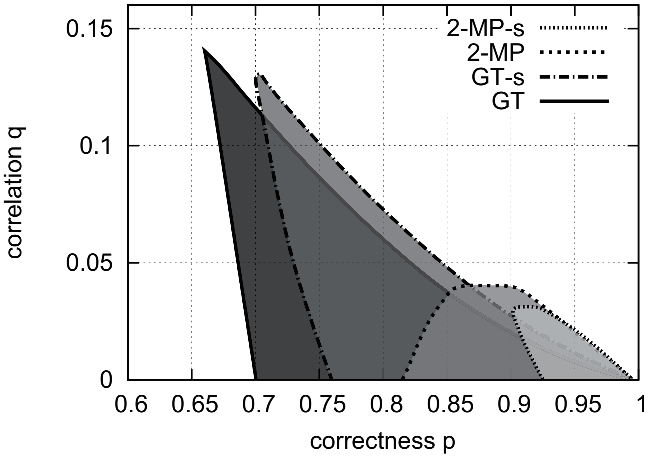

Fig. 14. Ranges of signal parameters over which GT/2-MP and GT-s/2-MP-s are SPFSE.

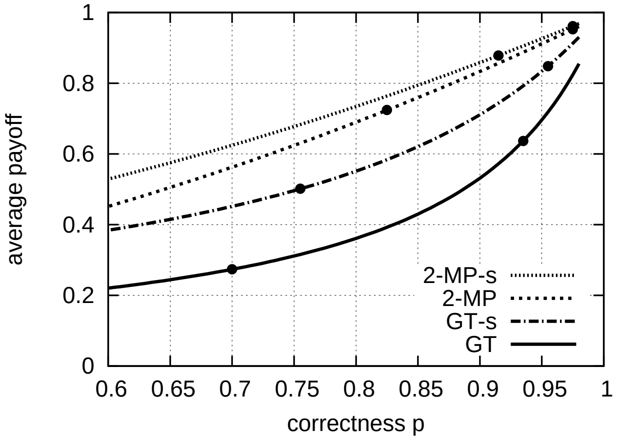

Fig. 15. Average payoff per period of GT/2-MP and GT-s/2-MP-s (q=0.01).

Then, how can players utilize (or disregard) this signal? Let us assume a parameter setting where GT constitutes an SPFSE. Since this signal is totally independent from the utilities/observation of players, disregarding this signal never hurts. Thus, GT (which disregards the signal) still constitutes an equilibrium.

Now let us assume player 2 uses the following strategy: as long as the event is not observed, play GT, but when the event is observed, move to state R. Then, assuming player 2 uses this strategy, for player 1, using the same strategy as player 2 would be a best response. This is because if player 1 observes the event, it is very likely that player 2 also observes the event and moves to state R. As long as the probability that player 2 is in state R is high, the best response for player 1 is to move to state R, since GT constitutes an SPFSE. Thus, this new strategy, which we call GT-s, can constitute an SPFSE. We call a similar modification of k-MP as k-MP-s. In summary, such a signal can serve as a “reset button” to restart a new repeated game, which makes punishments less severe.

We examine the range of parameters where GT-s or 2-MP-s constitutes an SPFSE, where $p' = 0.88, s' = 0.1$, and $(1 - p' - s')/2 = 0.01$. Figure 14 illustrates the ranges of signal parameters over which GT/2-MP and GT-s/2-MP-s are SPFSE. We can see that the range of GT-s (2-MP-s) is smaller than that of GT (2-MP) for the probability p that signals are correct for both players. On the other hand, only for GT, the range is larger for the probability q that either player observes the wrong signal. Figure 15 illustrates the average payoffs per period. We can see that the range over which GT-s (2-MP-s) is an SPFSE is smaller than that of GT (2-MP). However, the average payoffs still increase by introducing the additional signals.

A similar idea is presented in [Ellison 1994], but in that work, the signal is assumed to be public. By utilizing a POMDP solver, we can analyze the case where the signal is almost public.

7. CONCLUSION

This paper investigates repeated games with imperfect private monitoring. Although analyzing such games has been considered as a hard problem, we develop a program that automatically checks whether a given profile of FSAs can constitute an SPFSE. Our program is based on the ideas presented in our previous work [Kandori and Obara 2010] and utilizes an existing POMDP solver. This program enables non-experts of the POMDP literature, including researchers in the game theory, AI, and agent research communities, to analyze the equilibria of repeated games.

Furthermore, as a case study to confirm the usability of this program, we identify equilibria in an infinitely repeated prisoner's dilemma game with imperfect private monitoring, where the probability of an error is relatively small. We first examine how observation errors affect the behavior of GT, TFT, and 1-MP (Pavlov). Then we propose the k-MP strategy, which incorporates GT and Pavlov as a special case, and show that k-MP constitutes an SPFSE in a reasonably wide range of observation errors. Its efficiency is better than that of 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 nor is more efficient than GT. In our future work, we hope to investigate other games, such as congestion games, which can model various application problems including a packet routing problem, by utilizing our program.

REFERENCES

DOSHI, P. AND GMYTRASIEWICZ, P. J. 2006. On the Difficulty of Achieving Equilibrium in Interactive POMDPs. In *AAAI*. 1131–1136.

ELLISON, G. 1994. Cooperation in the Prisoner's Dilemma with Anonymous Random Matching. *Review of Economic Studies 61,* 3, 567–88.

FUDENBERG, D. AND TIROLE, J. 1991. *Game theory*. MIT Press.

HANSEN, E. A., BERNSTEIN, D. S., AND ZILBERSTEIN, S. 2004. Dynamic programming for partially observable stochastic games. In *AAAI*. 709–715.

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, M. 2010. *Repeated Games*. Game Theory, Palgrave macmillan, 286–299.

KANDORI, M. AND OBARA, I. 2010. Towards a Belief-Based Theory of Repeated Games with Private Monitoring: An Application of POMDP. mimeo.

KRAINES, D. AND KRAINES, V. 1989. Pavlov and the prisoner's dilemma. *Theory and Decision 26*, 47–79.

MAILATH, G. AND SAMUELSON, L. 2006. *Repeated Games and Reputation*. Oxford University Press.

MARECKI, J., GUPTA, T., VARAKANTHAM, P., TAMBE, M., AND YOKOO, M. 2008. Not All Agents Are Equal: Scaling up Distributed POMDPs for Agent Networks. In *AAMAS*. 485–492.

NAIR, R., TAMBE, M., YOKOO, M., PYNADATH, D., AND MARSELLA, S. 2003. Taming decentralized POMDPs: Towards efficient policy computation for multiagent settings. In *IJCAI*. 705–711.

NG, S.-K. AND SEAH, W. K. G. 2008. Game-Theoretic Model for Collaborative Protocols in Selfish, Tariff-Free, Multihop Wireless Networks. In *27th Conference on Computer Communications*. 216–220.

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.

PHELAN, C. AND SKRZYPACZ, A. 2012. Beliefs and Private Monitoring. *Review of Economic Studies*. to appear.

RAPOPORT, A. AND CHAMMAH, A. 1965. *Prisoner's Dilemma*. University of Michigan Press.

ROGERS, A., DASH, R. K., RAMCHURN, S. D., VYTELINGUM, P., AND JENNINGS, N. R. 2007. Coordinating team players within a noisy iterated Prisoner's Dilemma tournament. *Theoretical Computer Science 377,* 1–3, 243–259.

RUSSELL, S. AND NORVIG, P. 2009. *Artificial Intelligence: A Modern Approach (3rd Edition)*. Prentice Hall.

TENNENHOLTZ, M. AND ZOHAR, A. 2009. Learning equilibria in repeated congestion games. In *AAMAS*. 233–240.

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.