Towards a Belief-Based Theory of Repeated Games with Private Monitoring: An Application of POMDP[*]

KANDORI, Michihiro [†]
Faculty of Economics
University of Tokyo

OBARA, Ichiro [‡]
Department of Economics
UCLA

This Version: June 4, 2010

Abstract

An equilibrium in a repeated game with imperfect private moitoring is called a *finite state equilibrium*, if each player's action *on the equilibrium path* is given by an automaton with a finite number of states. We provide a tractable general method to check the equilibrium conditions in this class. Our method is based on the belief-based approach and employs the theory of POMDP (Partially Observable Markov Decision Processes). This encompasses the majority of existing works.

1 Introduction

Repeated games with imperfect private monitoring represent long-term relationships, where each player receives a noisy private signal about others' actions. Although this class of games has a wide range of applications, the characterization of all equilibria has yet to be obtained. This is in sharp contrast to the theory of repeated games with *perfect* or *imperfect public* monitoring, where complete characterizations of all equilibria have been obtained. The present paper provides valuable general methods to verify equilibrium conditions in repeated games with private monitoring.

In particular, we focus on a *finite state equilibrium*, where each player's action *on the equilibrium path* is given by an automaton with a finite state space. We provide a complete characterization of this class of equilibria and provide a tractable computational method to determine if a given profile of finite automata (one for

---

[*]Previous versions of this paper has been circulated under the title "Finite State Equilibria".

[†]e-mail: kandori@e.u-tokyo.ac.jp

[‡]e-mail: iobara@econ.ucla.edu

each player to determine his action on the path of play) can constitute a (finite) equilibrium. The present paper provides a unifying general theory to encompass the majority of the existing works, because most of them are based on some form of finite state equilibria. The belief-based approach by Sekiguchi (1997), Bhaskar and Obara (2002) consider trigger strategies on the path of play, hence finite state equilibria. The present paper can be regarded as a generalization of those papers. The belief-free approach by Ely and Välimäki (2002) considers an equilibrium which can be implemented by a finite state automata on and off the path of play. Proposition 4 in Ely, Horner and Olszewski (2005) (bang-bang property) shows that most of belief-free equilibrium payoffs (indeed all of them if the discount factor is close to 1) can be obtained by a finite state equilibrium with only two states.

Matsushima's review strategy equilibria (2004) and Hörner and Olszewski (2006) also employ finite equilibria. An exception which does not employ a finite equilibrium is Piccione (2002), whose equilibrium path requires infinite (countably many) states. However, the result by Ely, Horner and Olszewski shows that Piccione's equilibrium payoff can be obtained by a finite state equilibrium.

A more recent paper by Phelan and Skrzypacz (2009) also proposes an algorithm to compute a class of stationary finite state equilibria. Their approach focuses on dynamics of beliefs, while we utilize the theory of POMDP (Partially Observable Markov Decision Processes) to solve the belief-based dynamic programming problem.

2   Repeated Games with Private Monitoring

We use repeated games with private monitoring as our base model. Let A_i be the (finite) set of actions for player $i = 1, ..., N$ and $A := A_1 \times \cdots \times A_N$. Within each period, player i observes her own action a_i and private signal $\omega_i \in \Omega_i$. We denote $\omega = (\omega_1, ..., \omega_N) \in \Omega := \Omega_1 \times \cdots \times \Omega_N$ and let $q(\omega|a)$ be the probability of private signal profile ω given action profile a (we assume that Ω is a finite set). We denote the marginal distribution of ω_i by $q_i(\omega_i|a)$. It is also assumed that no player can infer which actions were taken (or not taken) for sure; to this end, we assume that each $\omega \in \Omega$ occurs with a positive probability for any $a \in A$ (*full support assumption*). 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

$$g_i(a) = \sum_{\omega \in \Omega} \pi_i(a_i, \omega_i) q(\omega|a).$$

This formulation ensures that the realized payoff π_i conveys no more information than a_i and ω_i do. A mixed action for player i is denoted by $\alpha_i \in \Delta(A_i)$, where $\Delta(A_i)$ is the set of probability distributions over A_i. With an abuse of notation, we denote the expected payoff and the signal distribution under a mixed action profile $\alpha = (\alpha_1, ..., \alpha_N)$ by $g_i(\alpha)$ and $q(\omega|\alpha)$ respectively. The stage game is to be played

repeatedly over an infinite time horizon $t = 1, 2, ...$ Player i's discounted payoff from a sequence of action profiles $a(t) \in A, t = 0, 1, 2...$ is given by $\sum_{t=1}^{\infty} \delta^t g_i(a(t))$, where $\delta \in (0, 1)$ is the discount factor.

3 Repeated Game Strategies and Path 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.

3.1 Repeated Game Strategies

A *private history* for player i at the beginning of time t is the record of player i's past actions and signals, $h_i^t = (a_i(0), \omega_i(0), ..., a_i(t-1), \omega_i(t-1)) \in H_i^t := (A_i \times \Omega_i)^t$. 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: formally, $s_i : H_i \to A_i$, where $H_i = \cup_{t \geq 0} H_i^t$. Similarly, a (behaviorally) mixed strategy for player i is denoted by $\sigma_i : H_i \to \Delta(A_i)$.

3.2 Path Automata and Finite State Equilibria

A **path automaton** $M_i \equiv (\Theta_i, \widehat{\theta}_i, f_i, T_i)$ of player i specifies the path of play (but not the behavior off the path of play) for player i, by specifying the following:

1. a set of states Θ_i

2. the initial state $\widehat{\theta}_i \in \Theta_i$

3. (pure) action choice for each state, $f_i : \Theta_i \to A_i$ (without loss of generality, we can assume that a pure action is played in each state[1].)

4. (possibly stochastic) state transition $T_i : \Theta_i \times \Omega_i \to \Delta(\Theta_i)$. Specifically, $T_i(\theta_i(t+1)|\theta_i(t), \omega_i(t))$ is the probability of the next state being $\theta_i(t+1)$ given the current state $\theta_i(t)$, action $a_i(t) = f_i(\theta_i(t))$, and private signal $\omega_i(t)$.

A path automaton without the specification of the initial state, denoted by $m_i \equiv (\Theta_i, f_i, T_i)$, is referred to as a **path preautomaton**. This concept turns out to be useful in our analysis. For any path preautomaton m_i, we denote the corresponding path automaton with initial state $\theta_i \in \Theta_i$ by (m_i, θ_i).

---

[1] Stochastic transition function can represent mixed action. For example, suppose that action C and D are played with an equal probability at state θ. Then, we can split this state θ into two states θ^C and θ^D and assume that in state θ^a, pure action a is played $(a = C, D)$. Furthermore we can specify the stochastic state transision function such that, in the event that θ is to be the next stae, state θ^C or θ^D realizes with an equal probability.

An important part of our definition above is that the transition function T_i presumes that the equilibrium action $a_i(t) = f_i(\theta_i(t))$ is played. Hence, our path automaton does *not* specify the behavior after a disequilibrium action $a_i(t) \neq f_i(\theta_i(t))$ (and therefore a path automaton only represents a part of a repeated game strategy).

To represent a repeated game strategy, one need to extend the transition function to specify the state transition after a deviating action $a_i(t) \neq f_i(\theta_i(t))$. To this end, let us define an *extended transition function* $\overline{T}_i : \Theta_i \times \Omega_i \times A_i \to \Delta(\Theta_i)$, where $\overline{T}_i(\theta_i(t+1)|\theta_i(t), \omega_i(t), a_i(t))$ is the probability of the next state being $\theta_i(t+1)$ given the current state $\theta_i(t)$, an arbitrary current action $a_i(t)$, and the current private signal $\omega_i(t)$. An *extended automaton* is denoted by $\overline{M}_i \equiv (\Theta_i, \widehat{\theta}_i, f_i, \overline{T}_i)$. An extended automaton specifies an action after any private history, so it induces a full repeated game strategy. A message of the present paper is that, somewhat surprisingly, path automata often turn out to be more useful than extended automata in the belief-based analysis of repeated games with private monitoring. A *finite path automaton* is a path automaton with a finite number of states. A finite path preautomaton and a finite extended automaton are defined similarly.

We are interested in equilibria where each player's behavior on the equilibrium can be described by a finite path automaton. We call a profile of finite path preautomata $m = (m_1, ..., m_N)$ **compatible** if, for every i, there exists some state $\theta_i \in \Theta_i$ and some belief $b_i \in \Delta(\Theta_{-i})$ such that (m_i, θ_i) is the optimal plan given his subjective belief b_i. Compatibility is necessary for any profile of finite path preautomata to be played on the equilibrium path. Note that compatibility does not guarantee that a set of such state-belief pairs across players are **consistent**: it may not be generated by some joint distribution on Θ. Suppose that there is a common prior $r \in \Delta(\Theta) = \Delta(\Theta_1 \times \cdots \times \Theta_N)$ such that (m_i, θ_i) is optimal against $r(\cdot|\theta_i) \in \Delta(\Theta_{-i})$ for every $\theta_i \in \Theta_i$ in the support of r and $i = 1, .., N$. Since we assume full support (of the marginal distribution on Ω_i), such profile of finite path preautomata and a joint distribution r on Θ constitutes a (correlated) sequential equilibrium once optimal strategies are assigned after every off the equilibrium history.[2] A *finite state equilibrium* is such a correlated sequential equilibrium where on-path behavior can be represented by finite state preautomata and a joint distribution on the product state space.

Definition 1 *A **finite state equilibrium** is a (correlated) sequential equilibrium of a repeated game with private monitoring, where players' behavior on the equilibrium path is given by finite path preautomata $m_i \equiv (\Theta_i, f_i, T_i)$, $i = 1, ..., N$ and a joint probability distribution of the initial states $r \in \Delta(\Theta)$.*

---

[2]Since we assume full support, (1) there is no ambiguity regarding the definition of sequential equilibrium although our game is an infinite horizon game and (2) every correlated sequential equilibrium is a correlated equilibrium i.e. sequential rationality is not a restriction.

The probability distribution r is the *initial correlation device*. At the beginning of $t = 0$, a profile of recommended initial states $\widehat{\theta}$ realizes with probability $r(\widehat{\theta})$, and player i observes his recommended initial state $\widehat{\theta}_i$. In a finite state equilibrium, each player i finds it optimal to follow path automaton $(m_i, \widehat{\theta}_i)$, given that others obey their recommended initial states. In other words, (m, r) constitutes a correlated equilibrium. A standard argument shows that (m, r) can always be "extended" to obtain a sequential equilibrium:

Remark 1 *The path preautomata $m_1, ..., m_N$ do not specify how each player should behave after his own deviations, but we can always find, for each player i, a repeated game strategy $\sigma_i(m_i)$ such that (i) it specifies the same behavior as m_i on the equilibrium path and (ii) it specifies an optimal continuation strategy for each information set of player i.[3] We will refer both $(\sigma_1(m_1), ..., \sigma_N(m_N), r)$ and (m, r) as a finite state equilibrium, if no confusion ensues.*

Note that, in a finite state equilibrium, a player's behavior off the equilibrium path need not be described by a finite automata. In fact, there is an equilibrium identified in the existing literature which has the on-path behavior described by finite path automata, while specifying its off-path behavior requires an extended automaton with infinitely many states:

Example 1 *The following example is from Bhaskar and Obara (2002). Consider a standard repeated prisoner's dilemma game with almost perfect and independent private monitoring: $A_i = \{C, D\}, \Omega_i = \{c, d\}, \Pr(\omega_i = c | a_i, C) = \Pr(\omega_i = d | a_i, D) = 1 - \varepsilon$ for any $a_i \in A_i$. Suppose that player 2's play is determined by the following "trigger strategy" preautomaton m, $\Theta = \{R, P\}, f(R) = C, f(P) = D, T(R | R, c) = 1, T(P | R, d) = T(P | P, \omega_2) = 1$ for any $\omega_2 \in \Omega_2$. Let b_1 be player 1's belief that player 2 is at state R. For a certain range of discount factor and small enough $\varepsilon > 0$, it can be shown that the optimal strategy against this preautomaton can be represented as a simple cut-off strategy ("belief-based strategy") in terms of belief: play C if $b_1 > b^*$ and play D if $b_1 < b^*$ for some cut-off belief $b^* \in (0, 1)$. It can be shown that this optimal strategy can be described by a finite state automaton given any belief. In fact, it is exactly given by the above preautomaton itself: (m, P) is the optimal strategy for $b_1 \in [0, b^*]$ and (m, R) is the optimal strategy for $b_1 \in [b^*, \overline{b}]$*

---

[3] The strategy $\sigma_i(m_i)$ is obtained by specifying an optimal continuation strategy for each information set of player i that is not reached on the equilibrium path (such an information set can be reached only by player i's own deviations). This does not change player i's payoff, and it does not affect other playres' best replies either (because player i's deviation is never detected by other players under the full support assumption). It remains to check if $\sigma_i(m_i)$ specifies optimal continuation strategies on the information sets that are reached on the equilbrium path. If this were not true, player i's payoff associated with m_i could be improved by replacing the supoptimal continuation strategy (specified by m_i) with an optimal one in some information set that is reached with a positive probability. This would imply that m_i were not a best reply, contradicting our premise that (m, r) is a correlated equilibrium.

$([0, \overline{b}]$ *is a belief-closed set, i.e. player 1's belief never leaves* $[0, \overline{b}]$ *when the initial belief is in* $[0, \overline{b}]$.. *So we don't have to consider* b_1 *above* $\overline{b}$).

However, player 1's optimal strategy off the equilibrium path cannot be described by a finite state automaton. Consider a history where b_1 is close to 0, for example, a history such as $(..., Dd, Dd, Dd, Dd)$. Suppose that player 1 deviates repeatedly by playing C (because D is the unique optimal action for $b_1 \in [0, b^*]$) and observes c in many periods. In this off the equilibrium path, player 1's belief eventually exceeds b^* after many realization of (Cc), after which the optimal continuation strategy must be (m, R). To implement such a strategy off the equilibrium path by an extended finite state automaton, one needs to use as many number of states as the required number of Cc to go back to (m, R). But this number can be arbitrarily large if b_1 is arbitrarily close to 0. Hence there is no finite extended automaton to implement such a strategy..

4   Private Monitoring Problem as Partially Observable Markov Decision Process (POMDP).

In a finite state equilibrium, player i' opponents' equilibrium behavior is given by finite path preautomata $m_j = (\Theta_j, f_j, T_j)$, $j \neq i$.   Under the full support assumption of private signals, player i never receives an evidence that his opponents $j \neq i$ deviated from their equilibrium behavior, so that he always believes that his opponents are using their path preautomata.   Hence, after any private history, player i's information regarding the current and future behavior of player $j \neq i$ is summarized by his belief over his opponents' states $b_i \in \triangle(\Theta_{-i})$.   Belief b_i compactly summarizes the relevant information contained in a private history $h_i^t = (a_i(1), \omega_i(1), ..., a_i(t), \omega_i(t))$ (a much more complicated object than b_i).

Furthermore, player i's belief at time t, denoted by $b_i(t)$, can be calculated by his previous belief $b_i(t-1)$, action $a_i(t-1)$ and signal $\omega_i(t-1)$ (the dynamics of beliefs is a controlled Markov process). This means that, *in a finite state equilibrium, each player i faces a relatively simple problem: a Markov decision problem (dynamic programming problem) with a finite dimensional state space* $\triangle(\Theta_{-i})$. This point has been emphasized by, for example, Bhaskar and Obara (2002), Mailath and Morris (2002), and Phelan and Skrzypacz (2009).   We go one step further and observe that this problem corresponds to a reasonably tractable class of Markov decision problems, known as POMDP (Partially Observable Markov Decision Process). The present paper introduces a general technique to solve such a Markov decision problem, building heavily on a technique developed in operations research and computer science (see, for example, Kaelbling, Littman, and Cassandra (1998)).

A crucial observation is that this decision problem is easier to solve than it appears. An apparent difficulty in solving this problem is that the value function W_i is defined on *uncountable number of states* $b_i \in \triangle(\Theta_{-i})$.   Hence, computing value iteration $W_i^{n+1}(b_i) = \Gamma_i W_i^n(b_i)$ or verifying the fixed point equation $W_i^*(b_i) =$

$\Gamma_i W_i^*(b_i)$ *for all* b_i in principle involves uncountably many calculations (one for each b_i). This is certainly true for a general non-linear value functions. Fortunately, however, the theory of POMDP shows that we can confine attention to *piecewise linear value functions*, and for those particular value functions the computation can be exactly done with a finite number of calculations.

Let us explain how the theory of POMDP works to verify that a given finite path preautomata $m_i \equiv (\Theta_i, f_i, T_i)$, $i = 1, ..., N$, constitute a correlated sequential equilibrium for some initial correlation device $r \in \Delta(\Theta)$. We will focus on player i's decision problem, given the opponents' finite state path preautomata m_{-i}. The relevant state for player i is the profile of his opponents' state θ_{-i}, which evolves by the Markov transition functions T_j ($j \neq i$). However, the true state is not directly observable, and player i only obtains partial information through his private signal ω_i. This is an instance of a Markov decision process with partially observable state variable.

The joint distribution of current signal ω_i and the next state θ'_{-i} given the current state and action (θ_{-i}, a_i) is given by

$$r_i(\omega_i, \theta'_{-i} | \theta_{-i}, a_i) \equiv \sum_{\omega_{-i}} \prod_{j \neq i} T_j^*(\theta'_j | \theta_j, \omega_j) q(\omega_i, \omega_{-i} | a_i, f_{-i}(\theta_{-i})), \qquad (1)$$

Using r_i thus defined, we can derive player i's posterior belief $\chi_i[a_i, \omega_i, b_i]$ after current action a_i and private signal ω_i given her current belief $b_i \in \Delta(\Theta_{-i})$:

$$\begin{aligned} \chi_i[a_i, \omega_i, b_i](\theta'_{-i}) &= \frac{\Pr^{b_i, a_i}(\omega_i, \theta'_{-i},)}{\Pr^{b_i, a_i}(\omega_i)} \\ &= \frac{\sum_{\theta_{-i}} r_i(\omega_i, \theta'_{-i} | \theta_{-i}, a_i) b_i(\theta_{-i})}{\sum_{\theta_{-i}} q_i(\omega_i | a_i, f_{-i}(\theta_{-i})) b_i(\theta_{-i})}. \end{aligned}$$

Given any value function $W_i : \Delta(\Theta_{-i}) \to \Re$, define the *Bellman operator* $\Gamma_i W_i$ by

$$\Gamma_i W_i(b_i) = \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (2)$$

$$\max_{a_i \in A_i} \sum_{\theta_{-j}} b_i(\theta_{-i}) \left[\sum_{\theta_{-j}} g_i(a_i, f_{-i}(\theta_{-i})) + \delta \sum_{\omega_i} W_i(\chi_i[a_i, \omega_i, b_i]) q_i(\omega_i | a_i, f_{-i}(\theta_{-i})) \right].$$

Call $W_i(b_i)$ a *belief-based value function*. The standard theory of dynamic programming shows that the optimal value function $W_i^*(b_i)$, which represents the maximum discounted payoff to player i under current belief b_i, is the unique fixed point $\Gamma_i W_i^* = W_i^*$. Furthermore, W_i^* can be obtained as the limit of the sequence of value functions $\{W_i^n\}$, where $W_i^{n+1} = \Gamma_i W_i^n$ (W_i^0 can be any function).

The theory of POMDP shows that the Bellman operator $\Gamma_i W_i$ (2) has a much simpler expression for some W_i. To explain how it works, we need to introduce a couple of concepts. First, for any finite set of player i's path automata $\mathcal{M}_i = \{M_i^1, ..., M_i^K\}$, consider a path automaton such that

1. it starts with a state where some pure action a_i is played and

2. after ω_i is realized, an automaton in $\mathcal{M}_i$, denoted by $M_i(\omega_i)$, is implemented.

Such a path automaton $(a_i, M_i(\cdot))$ is called a **one-shot extension of path automata** $\mathcal{M}_i$. It is a path automaton constructed by attaching an initial state to the set of path automata $\mathcal{M}_i$. Let $\widetilde{\mathcal{M}}_i$ denote the set of all one-shot extensions of $\mathcal{M}_i$. Note that $\widetilde{\mathcal{M}}_i$ has a finite number ($= |A_i||\mathcal{M}_i|^{|\Omega_i|}$) of elements.

Second, for any path automaton M_i of player i, let $v_i^{M_i\theta_{-i}}$ be player i's payoff associated with M_i, when the opponents' states are θ_{-i}. Then the expected payoff associated with M_i under belief $b_i \in \triangle(\Theta_{-i})$ is given by

$$V_i^{M_i}(b_i) = \sum_{\theta_{-i}} v_i^{M_i\theta_{-i}} b_i(\theta_{-i}). \tag{3}$$

As an expected payoff, this function is linear in belief b_i. This fact plays an important role in what follows.

Given those concepts, the essence of POMDP can be summarized by the following proposition. It shows that the Bellman operator $\Gamma_i W_i$ (2) has a very simple representation for some particular value functions W_i:

Proposition 1 (POMDP) *Let $\mathcal{M}_i$ be a set of player i's path automata and consider value function $W_i(b_i) \equiv \max_{M_i \in \mathcal{M}_i} V^{M_i}(b_i)$. Then, the Bellman operator for this value function is given by*

$$\Gamma_i W_i(b_i) = \max_{M_i \in \widetilde{\mathcal{M}}_i} V_i^{M_i}(b_i),$$

where $\widetilde{\mathcal{M}}_i$ is the set of one-shot extensions of $\mathcal{M}_i$.

The proof is given by a direct calculation. Here we provide an intuition. The value function $W_i(b_i) \equiv \max_{M_i \in \mathcal{M}_i} V_i^{M_i}(b_i)$ is simply player i's payoff under the constrained best reply to b_i (given that player i must chose a path automata in set $\mathcal{M}_i$). Hence, the maximand on the right-hand side of the Bellman operator $\Gamma_i W_i$ (2) is the payoff associated with some action a_i today, followed by a best path automata in $\mathcal{M}_i$. Hence, $\Gamma_i W_i(b_i)$ is the payoff associated with the best one-shot extension of $\mathcal{M}_i$, given belief b_i ($= \max_{M_i \in \widetilde{\mathcal{M}}_i} V_i^{M_i}(b_i)$).

Why is this proposition useful? In principle, computing $\Gamma_i W_i(b_i)$ for all $b_i \in \triangle(\Theta_{-i})$ involves uncountably many calculations (one for each b_i). However, computation of $\max_{M_i \in \widetilde{\mathcal{M}}_i} V^{M_i}(b_i)$ is easy. This is simply the upper envelope of a finite number of linear functions $V^{M_i}(b_i)$, $M_i \in \widetilde{\mathcal{M}}_i$, and it can be exactly calculated in a finite number of steps.

Now let us formally explain how the theory of POMDP works. Recall that our goal is to verify that the given finite path preautomaton profile $m = (m_1, ..., m_N)$

can constitute a finite state equilibrium. The POMDP procedure is describes as follows.

<u>*POMDP Value Iteration*</u>:

- Define the initial candidate path automata as $\mathcal{M}_i^0 \equiv \{(m_i, \theta_i)\}_{\theta_i \in \Theta_i}$.

- In step n (n=1,2...), given the set of initial candidate path automata $\mathcal{M}_i^{n-1}$, compute the value function by

$$W_i^n(b_i) = \max_{M_i \in \widetilde{\mathcal{M}}_i^{n-1}} V_i^{M_i}(b_i),$$

  where $\widetilde{\mathcal{M}}_i^{n-1}$ is the set of one-shot extensions of path automata $\mathcal{M}^{n-1}$. This can be computed in a finite number of steps, because W_i^n is the upper envelope of a finite number of linear functions $V_i^{M_i}(b_i)$. Then, construct the set of candidate path automata for the nest step by "cream skimming" the one-shot extensions:

$$\mathcal{M}_i^n \equiv \left\{ M_i \in \widetilde{\mathcal{M}}_i^{n-1} | W_i^n(b_i) = V_i^{M_i}(b_i) \text{ for some } b_i \right\}.$$

- This defines an increasing sequence of value functions $W_i^0 \leq W_i^1 \leq \cdots$, where, for each $n < \infty$, W_i^n is a piecewise linear, convex and continuous function (because it is the upper envelope of a finite number of linear functions). The optimal value function is obtained as the limit $W_i^* = \lim_{n \to \infty} W_i^n$.

The monotonicity of the sequence of value function is shown by a standard argument. It is easy to show $W_i^0 \leq W_i^1$.[4] By the definition, operator Γ_i is monotone: $W^1 \geq W^0$ implies $W^2 = \Gamma_i W_i^1 \geq \Gamma_i W_i^0 = W_i^1$. Proceeding inductively, we obtain $W^0 \leq W^1 \leq \dots$. Also the following is a standard result in POMDP.

Lemma 1 *The optimal value function W_i^* is convex and continuous.*

---

[4] The proof $W_i^0 \leq W_i^1$: If the transition functions of candidate automata in $\mathcal{M}_i^0$ are deterministic, the proof is trivial (because $\mathcal{M}_i^0 \subset \widetilde{\mathcal{M}}_i^0$). Here we consider the case with stochastic transition functions. $W_i^1(b_i)$ is the payoff when (i) an optimal action a_i^* is played today and (ii) the continuation play after ω_i is given by a best on-path automaton in $\mathcal{M}^0$, given the posterior belief $b_i = \chi_i[a_i^*, \omega_i, b_i]$. On the other hand, $W_i^0(b_i)$'s continuation payoff is asspciated with some on-path automataon in $\mathcal{M}^0$. Hence, it is the payoff when (i) some action is played today and (ii) the continuation play after ω_i is given by a probability distribution over $\mathcal{M}^0$ (this is given by the possibly stochastic transition function T_i of preoutomaton m_i). Since the choice of continuation automata in (ii) may not be optimal, $W_i^0(b_i)$ is no greater than $W_i^1(b_i)$, which uses optimal continuation automata in $\mathcal{M}^0$. Q.E.D.

Proof. Let $\mathcal{M}_i^*$ be the set of optimal path automata (i.e., an element of $\mathcal{M}_i^*$ is a path automaton which is optimal for some belief b_i). Then, $W_i^*(b_i)$ is the upper envelope of a family of linear functions $V_i^{M_i}(b_i)$, $M_i \in \mathcal{M}_i^*$. Hence it is convex and continuous. ∎

Remark 2 *An important goal in the literature on POMDP is to design a fast computer algorithm to implement the value iteration described above, and there are some free programs ("POMDP solvers") available over the internet.*

The POMDP value iteration provides a iterative procedure to find player i's best replies against m_{-j}, which represents various continuation strategies the opponents might use in the equilibrium. In each step n, a finite number of path automata $\mathcal{M}_i^n$ is given and the procedure search for better replies in a neighborhood of $\mathcal{M}_i^n$. This neighborhood (denoted $\widetilde{\mathcal{M}^n}_i$) is the set of one-shot extensions of $\mathcal{M}_i^n$. The cream skimming procedure is just to find undominated strategies in $\widetilde{\mathcal{M}^n}_i$ (in the game with restricted strategy spaces given by $\widetilde{\mathcal{M}^n}_i$ and m_{-i}). (Note that the cream skimming procedure is to find strategies which can be a best reply to some joint distribution over the opponents' states. A standard result in game theory shows that such a strategy is equivalent to a strategy that is not strictly dominated.) The POMDP iteration shows that *repeated search for better replies by finding undominated one-shot extensions* eventually leads to the best response.

A particularly useful implication is that verifying optimality is easy. To show that a finite set of path automata $\mathcal{M}_i^*$ provides best replies to m_{-i}, one only needs to show that there is no better replies in the one-shot extensions of $\mathcal{M}_i^*$. This is a counterpart of the celebrated one-shot deviation principle in the repeated games with public monitoring. We elaborate on this point in Section 5.1.

5   Finite State Equilibrium as Multiperson POMDP

So when does a given finite state path automaton $m = (m_1, ..., m_N)$ constitute a finite state equilibrium? Observe that (m_i, θ_i) is optimal given $b_i \in \triangle(\Theta_{-i})$ if and only if the discounted value generated by path automaton (m_i, θ_i) is exactly equal to the optimal value function $W_i^*(b_i)$. Hence m is compatible if and only if $W_i^*(b_i) = W_i^0(b_i)$ for some b_i for every i.

Proposition 2 *A profile of path preautomata $m = (m_1, ..., m_N)$ is compatible if there exists $b_i \in \Delta(\Theta_{-i})$ such that $W_i^*(b_i) = W_i^0(b_i)$ for $i = 1, ..., N$.*

Once we verify that m is compatible, then we need to find a joint distribution r on $\Delta(\Theta)$ such that (m, r) constitutes a correlated sequential equilibrium. Thus our basic computational procedure works as follows.

Verification Procedure:

Given finite path preautomata $m = (m_1, ..., m_N)$

1. Use the theory of POMDP to find the optimal value function W_i^*.

2. Look for the region of beliefs where the optimal value coincides with the payoff associated with a candidate path automaton (m_i, θ_i):

$$B_i^{\theta_i} \equiv \{b_i | W_i^*(b_i) = V_i^{(m_i, \theta_i)}(b_i)\}.$$

Since W_i^* is convex and continuous and $V_i^{(m_i, \theta_i)}$ is linear, $B_i^{\theta_i}$ is a (possibly empty) closed and convex set. If this is empty for all θ_i, m cannot constitute an equilibrium. Otherwise, proceed to 3.

3. Find an initial correlation device $r \in \Delta(\Theta)$ such that $r_i(\theta_i) > 0$ implies $r_{-i}(\cdot | \theta_i) \in B_i^{\theta_i} \neq \varnothing$ (r_i is the marginal distribution, and $r_{-i}(\cdot | \theta_i)$ is the conditional distribution of θ_{-i}). If there is such r, then (m, r) is a finite state equilibrium. If there is no such r, m cannot constitute an equilibrium.

We first discuss the first two steps, which concerns with the compatibility of path automata, then discuss the third step in the next subsection. To verify compatibility, we first need to find the optimal value function W_i^*. One possibility is that we find a fixed point $W_i^n = \Gamma_i W_i^n = W_i^*$ in a finite number of step. We have examples of this case in Sections 6.1 - 6.4.

It is sometimes useful to focus on a subset of beliefs. A *belief-closed set* $X_i \in \Delta(\Theta_{-i})$ for player i is the set of player i's beliefs such that player i's posterior belief will never leave it if player i's current belief is in it after any private history of player i (including off the equilibrium history). Formally, X_i is a belief-closed set if $\chi_i[a_i, \omega_i, b_i] \in X_i$ for any $a_i \in A_i, \omega_i \in \Omega_i$ and $b_i \in \Theta$. If we can find a fixed point $W_i = \Gamma_i W_i$ on X_i, then it must be the case that $W_i(b_i) = W_i^*(b_i)$ on X_i. This is because player i's dynamic programming problem is unchanged even if we use a smaller state space X_i when the initial belief b_i is in X_i. Examples of this case are provided in Sections 6.1- 6.3. The previous case is a special case where $X_i = \Delta(\Theta_{-i})$, which is obviously a belief-closed set. The following proposition summarizes this observation.

Proposition 3 *A profile of path preautomata* $m = (m_1, ..., m_N)$ *is compatible if there exists a belief-closed set* $X_i \in \Delta(\Theta_{-i})$ *and* W_i *for each* i *such that (1)* $W_i(b_i) = \Gamma_i W_i(b_i)$ *for every* $b_i \in X_i$ *and (2)* $W_i(b_i') = V_i^{(m_i, \theta_i)}(b_i')$ *for some* $b_i' \in X_i$ *and some* θ_i.

Another possibility is that W_i^n is strictly increasing in each step (at least for some beliefs), hence it takes (or seems to take) infinite iterations to reach the optimal value function. Now suppose that, in such a case, there is a finite n and some belief b_i where $W_i^n(b_i) = V_i^{(m_i,\theta_i)}(b_i)$ $(= W_i^0(b_i))$. Even if this holds for a fairly large n, however, there is a possibility that in a succeeding step $k > n$ the value function shifts upward $W_i^k(b_i) > V_i^{(m_i,\theta_i)}(b_i)$ and the optimality of automaton (m_i, θ_i) at belief b_i is disproved. *How can we verify the optimality of a candidate automaton in a finite step*, when W_i^n does not (seem to) converge to W_i^* in any finite step? The following proposition addresses this issue.

To state our proposition, we need to define the following concepts. Given a profile of on-path preautomata m, we define a **profile of on-path belief closed sets** for player i as

$$X_i = (X_i(\theta_i))_{\theta_i \in \Theta_i},$$

that satisfies $\forall \theta_i$ $X_i(\theta_i) \subset \triangle(\Theta_{-i})$ ($X_i(\theta_i)$ can be an empty set) and the following property: For any $\theta_i', \theta_i'', b_i, \omega_i,$

$$\text{if } b_i \in X_i(\theta_i') \text{ and } T_i(\theta_i''|\theta_i', \omega_i) > 0, \text{ then } \chi_i[f_i(\theta_i'), \omega_i, b_i] \in X_i(\theta_i'').$$

This means that player i's posterior belief never leaves the on-path belief closed sets as long as player i does not deviate from the specified action at each state. Let $p(M_i, X_i, b_i)$ be the probability that the posterior belief in the next period moves outside of X_i when the current belief is b_i and a path-automaton M_i is played. For any belief-based value functions V and W, define $|V - W| \equiv \sup_{b_i \in \triangle(\Theta_{-i})} |V(b_i) - W(b_i)|$.

Proposition 4 (*Optimality Verification in A Finite Step*) *Let $X_i = (X_i(\theta_i))_{\theta_i \in \Theta_i}$ be any profile of on-path belief closed sets of player i with respect to m. Suppose that, at the nth step of the POMDP value iteration, for any θ_i and $b_i \in X_i(\theta_i)$, we have $W_i^n(b_i) = V_i^{(m_i,\theta_i)}(b_i)$ and*

$$V_i^{(m_i,\theta_i)}(b_i) \geq \max_{M_i \in \widetilde{\mathcal{M}}_i^n} \left\{ V_i^{M_i}(b_i) + p(M_i, X_i, b_i) \frac{\delta^{n+1}}{1-\delta} |W_i^1 - W_i^0| \right\}. \tag{4}$$

Then, path automaton (m_i, θ_i) is optimal at all $b_i \in X_i(\theta_i)$: $W_i^*(b_i) = V_i^{(m_i,\theta_i)}(b_i)$ for any $b_i \in X_i(\theta_i)$ and any $\theta_i \in \Theta_i$.

Remark 3 *This proposition is useful when (i) $W_i^n = V_i^{(m_i,\theta_i)}$ for some beliefs but (ii) for some other beliefs the value iteration takes (or seems to take) infinite steps to reach the optimal value function W_i^*. The intuition of this proposition is quite simple. The standard theory of dynamic programming shows that the value function at the nth iteration is very close to the optimal one, for a large n. In particular, the optimal value function is bounded above by $W_i^n + \frac{\delta^n}{1-\delta} |W_i^1 - W_i^0|$, and the condition (4) basically says that any one-shot deviation is unprofitable, even when the player can receive this upper bound of the optimal value off the equilibrium path.*

Proof. Define U_i^0 by $U_i^0(b_i) = V_i^{(m_i,\theta_i)}(b_i)$ if $b_i \in X_i(\theta_i)$ for some θ_i and $U_i^0(b_i) = W_i^*(b_i)$ for every other belief (outside of X_i). ($U_i^0(b_i)$ is well-defined even when b_i belongs to more than one sets $X_i(\theta_i), X_i(\theta_i'), \ldots$, because our premise requires $V_i^{(m_i,\theta_i)}(b_i) = V_i^{(m_i,\theta_i')}(b_i) = \cdots = W_i^n(b_i)$.) We will show that $U_i^0 = W_i^*$.

We first show that $U_i^0(b_i) = \Gamma_i U_i^0(b_i)$ if $b_i \in X_i(\theta_i)$ for some θ_i. When we denote player i's current expected stage payoff given current action a_i and belief b_i by $u_i(a_i, b_i)$ and the posterior belief in the next period by b_i', $\Gamma_i U_i^0(b_i)$ is expressed as

$$
\begin{aligned}
\Gamma_i U_i^0(b_i) &= \max_{a_i \in A_i} \left\{ u_i(a_i, b_i) + \delta E[U_i^0(b_i')|a_i, b_i] \right\} \\
&= u_i(a_i^1, b_i) + \delta E[U_i^0(b_i')|a_i^1, b_i].
\end{aligned}
$$

Since U_i^0 can be expressed as

$$
U_i^0(b_i') = \begin{cases} W_i^n(b_i') \; (= V_i^{(m_i,\theta_i)}(b_i')) \text{ if } b_i' \in X_i(\theta_i) \text{ for some } \theta_i \\ W_i^*(b_i') \text{ otherwise} \end{cases},
$$

we have

$$
\begin{aligned}
\Gamma_i U_i^0(b_i) &= u_i(a_i^1, b_i) + \delta E[W_i^n(b_i') + \{U_i^0(b_i') - W_i^n(b_i')\} |a_i^1, b_i]. \\
&\leq u_i(a_i^1, b_i) + \delta E[W_i^n(b_i')|a_i^1, b_i] + p\left(M_i', X_i, b_i\right)\delta|W_i^* - W_i^n|,
\end{aligned}
$$

where M_i' is any path automaton whose initial action is a_i^1. Recall that $p\left(M_i', X_i, b_i\right)$ is the probability that the posteriori b_i' moves out of the profile of belief-closed sets ($b_i' \notin X_i(\theta_i)$ for any θ_i) under automaton M_i', so that it only depends on the initial action of M_i'. Now, by the standard theory of dynamic programming ($|W_i^* - W_i^n| \leq \frac{\delta^n}{1-\delta}|W_i^1 - W_i^0|$), we have

$$
\Gamma_i U_i^0(b_i) \leq \max_{a_i} \left\{ u_i(a_i, b_i) + \delta E[W_i^n(b_i')|a_i, b_i] \right\} + p\left(M_i', X_i, b_i\right)\frac{\delta^{n+1}}{1-\delta}|W_i^1 - W_i^0|.
$$

Since the first term on the right hand side is equal to $\Gamma_i W_i^n = \max_{M_i \in \widetilde{\mathcal{M}}_i^n} V_i^{M_i}(b_i)$, we have

$$
\Gamma_i U_i^0(b_i) \leq \max_{M_i \in \widetilde{\mathcal{M}}_i^n} \left\{ V_i^{M_i}(b_i) + p\left(M_i, X_i, b_i\right)\frac{\delta^{n+1}}{1-\delta}|W_i^1 - W_i^0| \right\}. \tag{5}
$$

By the premise of the proposition (4), this is no greater than $V_i^{(m_i,\theta_i)}(b_i)$. Since $U_i^0(b_i) = V_i^{(m_i,\theta_i)}(b_i)$ for $b_i \in X_i(\theta_i)$, we have established that $\Gamma_i U_i^0(b_i) \leq U_i^0(b_i)$ if $b_i \in X_i(\theta_i)$ for some θ_i.

Now we show $\Gamma_i U_i^0(b_i) \geq U_i^0(b_i)$ if $b_i \in X_i(\theta_i)$ for some θ_i. This follows from the fact that, if $b_i \in X_i(\theta_i)$ for some θ_i,

$$
\Gamma_i U_i^0(b_i) \geq u_i(f_i(\theta_i), b_i) + \delta E[U_i^0(b_i')|f_i(\theta_i), b_i] = V_i^{(m_i,\theta_i)}(b_i) = U_i^0(b_i).
$$

13

The first equality holds because of the following reasoning. Suppose that player i plays action $f_i(\theta_i)$ today under current belief $b_i \in X_i(\theta_i)$, and consider the case where the posterior belief in the next period is $b_i' \in X_i(\theta_i')$. This means that tomorrow's state is θ_i' (for a moment, consider the case where (m_i, θ_i) has deterministic transition function). Therefore, the continuation payoff of automaton (m_i, θ_i) is $V_i^{(m_i, \theta_i')}(b_i')$. By definition, this is equal to $U_i^0(b_i')$ (because $b_i' \in X_i(\theta_i')$). Hence, $u_i(f_i(\theta_i), b_i) + \delta E[U_i^0(b_i') | f_i(\theta_i), b_i]$ represents the value associated with automaton (m_i, θ_i), so that it is equal to $V_i^{(m_i, \theta_i)}(b_i)$. The case of stochastic transition function is similar.[5] Hence, we have shown $U_i^0(b_i) = \Gamma_i U_i^0(b_i)$ for $b_i \in X_i(\theta_i)$, for some θ_i.

Lastly, for $b_i \notin X_i(\theta_i)$ for any θ_i, we show $\Gamma_i U_i^0(b_i) = U_i^0(b_i)$. Clearly $\Gamma_i U_i^0(b_i) \leq W_i^*(b_i) = U_i^0(b_i)$ if $b_i \notin X_i(\theta_i)$ for any θ_i, because $U_i^0(b_i) \leq W_i^*(b_i)$ for every b_i. Hence, we have $U_i^1(b_i) = \Gamma_i U_i^0(b_i) \leq U_i^0(b_i)$ for all b_i, and U_i^n would be a decreasing sequence and $\lim_{n\to\infty} U_i^n = W_i^*$ by the theory of dynamic programming. Therefore, if $U_i^1(b_i') = \Gamma_i U_i^0(b_i') < W_i^*(b_i') (= U_i^0(b_i))$ for some $b_i' \notin X_i(\theta_i)$ for any θ_i, we would have a contradiction $W_i^*(b_i') \leq U_i^1(b_i') < W_i^*(b_i')$. Hence, it must be the case that $\Gamma_i U_i^0(b_i) = U_i^0(b_i)$ if $b_i \notin X_i(\theta_i)$ for any θ_i.

Those arguments have shown that U_i^0 is a fixed point of Γ_i, so we obtain $W_i^* = U_i^0$ everywhere. In particular, $W_i^*(b_i) = U_i^0(b_i) = V_i^{(m_i, \theta_i)}(b_i)$ for any $b_i \in X_i(\theta_i)$ and any $\theta_i \in \Theta_i$. ∎

An example of Proposition 4 (Example 4) is provided in Section 6.5.

5.1 A Simpler Verification Method When Off-Path Candidate Automata Are Also Given

Suppose that you have some guess about "comprehensive" path preautomata m and a belief-closed set $X_i, i = 1, ..., N$ that work. That is, m should include not only the automata to be played on the equilibrium path but also all automata used off the equilibrium path. You expect that W_i^0, which is generated from m, is a fixed point on X_i. We can use POMDP to verify this numerically as described above. But there is a simpler, albeit equivalent, way to verify this.

When we have such a "comprehensive" candidate m, our verification procedure described above boils down to:

Belief-Based One-Shot Deviation Principle: Candidate preautomaton m_i provides the best replies against m_{-i} on a belief-closed set X_i, if it cannot be improved upon by one-shot extensions. That is, there is no one-shot extension M_i of

---

[5]When m_i has a stochastic transition function, two staes θ_i' and θ_i'' may be possible for a given posterior belief b_i' (because of the stochastic transition function). Hence the same posteior belief b_i' may be contained in $X_i(\theta_i')$ and $X_i(\theta_i'')$. However, the premise in this proposition guarantees $V_i^{(m_i, \theta_i')}(b_i') = V_i^{(m_i, \theta_i'')}(b_i') (= W_i^n(b_i'))$, so that we can employ the same argument as before.

$\{(m_i, \theta_i)|\theta_i \in \Theta_i\}$ and $b_i \in X_i$ such that $V_i^{M_i}(b_i) > V_i^{(m_i, \theta_i)}(b_i)$ for all θ_i.

This statement is equivalent to $\Gamma_i W_i^0(b_i) = W_i^0(b_i)$ for all $b_i \in X_i$, with $W_i^0(b_i) = \max_{\theta_i} V_i^{(m_i, \theta_i)}(b_i)$. In what follows, we show that, instead of checking $\Gamma_i W_i^0(b_i) = W_i^0(b_i)$ for all beliefs, we only need to check this at a finite number of beliefs (this is essentially because the maximized value function is the upper envelope of linear functions). Let $B_i^{\theta_i, 0} \subset \triangle(\Theta_{-i})$ be the set of beliefs where the following holds

$$W_i^0(b_i) = V^{(m_i, \theta_i)}(b_i).$$

Clearly $\bigcup_{\theta_i} B_i^{\theta_i, 0}$ covers $\triangle(\Theta_{-i})$. Since each $B_i^{\theta_i, 0}$ is an intersection of a finite number of half spaces (and a hyperplane $\sum_{\theta_{-i}} b_i(\theta_{-i}) = 1$), it is a closed convex polyhedron. Take any convex polyhedron D_i such that $X_i \subset D_i$. Define a finite subset of D_i as follows:

$$D_i^{\theta_i} = \left\{ b_i \in D_i | b_i \text{ is an extreme point of } D_i \cap B_i^{\theta_i, 0}, \theta_i \in \Theta_i \right\}.$$

Remember that the objective function of the dynamic programming problem is linear in beliefs. Hence, if we can verify that $\Gamma_i W_i^0(b_i) = W_i^0(b_i)$ for every $b_i \in D_i^{\theta_i}$, which is just a set of finite points, then we also obtain $\Gamma_i W_i^0(b_i) = W_i^0(b_i)$ every where in $D_i \cap B_i^{\theta_i, 0}$ i.e. W_i^0 is a fixed point on $D_i (\supset X_i)$.

Figure 1 illustrates our point. This corresponds to a two-player game, where the candidate path pre automaton of each player has three states, θ_i^1, θ_i^2, and θ_i^3. In the figure, $D_i \cap B_i^{\theta_i^k, 0}$ is denoted by B^k. D_i is the set of all beliefs ($= \triangle(\Theta_{-i})$) in this example, and player i has candidate optimal plans (m_i, θ_i^1), (m_i, θ_i^2), and (m_i, θ_i^3). $\bigcup_{\theta_i} D_i^{\theta_i}$ consists of seven points $b^1, ..., b^7$, and each of them can be easily computed. For example, b^3 is a solution to the system of linear equalities

$$\begin{cases} V^{(m_i, \theta_i^1)}(b_i) = V^{(m_i, \theta_i^2)}(b_i) \\ V^{(m_i, \theta_i^2)}(b_i) = V^{(m_i, \theta_i^3)}(b_i) \\ \sum_{\theta_{-i}} b(\theta_{-i}) = 1 \end{cases}.$$

Let $\Theta_i^{b_i} \subset \Theta_i$ be the set of states such that $B_i^{\theta_i, 0}$ includes $b_i' \in \bigcup_{\theta_i} D_i^{\theta_i}$. Our argument above can be summarized as follows:

Proposition 5 (Finite Verification of the Belief-Based One-Shot Deviation Principle) *To verify $\Gamma_i W_i^0 = W_i^0 = \max_{\theta_i} V_i^{(m_i, \theta_i)}$ on a belief-closed set X_i, it*
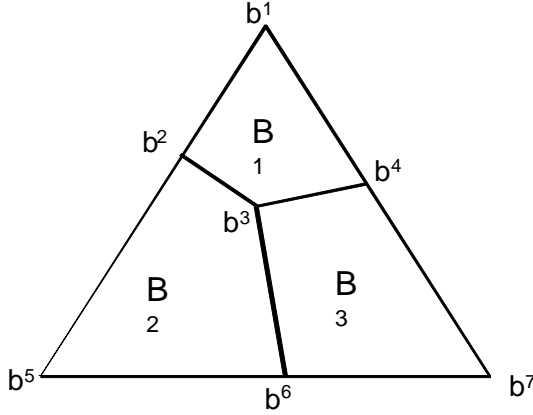
Figure 1:

is sufficient to check the following conditions ("no gain from one-shot deviations") at a finite number of "extreme point" beliefs $b_i \in \bigcup_{\theta_i} D_i^{\theta_i}$,

(1) for each $\theta_i \in \Theta_i^{b_i}$, path automaton (m_i, θ_i) transits to an optimal path automata among $\left\{ (m_i, \theta_i'), \theta_i' \in \Theta_i \right\}$ after every realization of private signal ω_i (because (m_i, θ_i) must be optimal given b_i), i.e. $\chi_i[f(\theta_i), \omega_i, b_i] \in B_i^{\theta_i', 0}$ for any ω_i and θ_i' with $T\left(\theta_i' | \theta_i, \omega_i \right) > 0$.

(2) for any action $a_i \in A_i$ that is not played by any $\theta_i \in \Theta_i^{b_i}$, no one-shot extension $(a_i, M_i(\cdot))$ should generate a larger value than $V^{(m_i, \theta_i)}(b_i)$ for any $\theta_i \in \Theta_i^{b_i}$.

Section 6.2 provides an example of this proposition.

5.2   When Can We Find a Right Initial Correlation Device?

We show that, when a profile of path preautomata is compatible, the existence of a right initial correlation device is guaranteed under a mild set of assumptions.

Theorem 4 *Let $m_i \equiv (\Theta_i, f_i, T_i)$, $i = 1, ..., N$ be the candidate finite path preautomata, and suppose that (i) Markov chain induced by $(m_1, ..., m_N)$ on Θ has a unique recurrent communication class[6] $X \subset \Theta$ (ii) X contains an aperiodic state[7]*

---

[6]A subset $X \subset \Theta$ is called a *recurrent communication class* if (i) any two points in X are mutually reachable and (ii) no point $\theta \notin X$ is reachable from any state within X.

[7]A state $x \in \Theta$ is *aperiodic* if the greatest common divisor of $\{t > 0 | \Pr(\theta(t) = x | \theta(0) = x) > 0\}$

and (iii) for all i, there is a belief b_i^0 such that $V^{\theta_i^0}(b_i^0) = W_i^*(b_i^0)$ for some $\theta_i^0 \in \Theta_i$, where W_i^* is the optimal value function. Then, there is a probability distribution μ over X such that choosing the initial states by μ and following $(m_1, ..., m_N)$ is a correlated equilibrium.

Example 2 *Let us present an example of conditions (i) and (ii). This is the example in the next section. Let m_i be the path preautomaton associated with the grim trigger strategy in our prisoner's dilemma game. The state space is $\Theta_i = \{R, P\}$, and R and P are the cooperative and non-cooperative states respectively. The Markov chain induced by (m_1, m_2) has a unique recurrent communication class $X = \{(P, P)\}$, and (P, P) is an aperiodic state (because $\{t > 0 | \Pr(\theta(t) = (R, R)|\theta(0) = (R, R)) > 0\} = \{1, 2, ...\}$).*

Remark: A similar observation is made in Phelan and Skrzypacz (2009). It seems that, like us, PS impose some sufficient conditions on the transition of Markov chain so that there is the unique and globally stable ergodic distribution. Since the joint distribution on Θ converges to the stationary distribution for any initial distribution, they can use the stationary distribution as the initial joint distribution to check if the finite extended automata to be an equilibrium. Then they check whether one-shot deviation constraint is satisfied at every possible subjective belief conditional on each state (but the set of conditional subjective beliefs is larger in their setting because they keep track of beliefs of on and off the equilibrium path). On the other hand, we first find a belief on which a given path automaton is optimal. Then we impose sufficient conditions for the existence of unique and globally stable ergodic distribution on Θ and verify that path preautomaton is optimal in the limit, i.e. when the ergodic distribution is used as a correlation device.

Proof. Let $MC(\Theta, m)$ be the Markov chain induced by $(m_1, ..., m_N)$ on Θ. The theory of Markov chain shows that there is a unique invariant distribution μ of $MC(\Theta, m)$ and (i) its support is X and (ii) for any initial distribution of state profile $\mu^0 \in \Delta(\Theta)$, the probability distribution of state profile at time t, denoted by μ^t, converges to μ.

By condition (iii) in the Theorem, each player i has a belief b_i^0 under which automaton (m_i, θ_i^0) is his optimal strategy. Let μ^0 be the initial distribution of Markov chain $MC(\Theta, m)$ where player i is in state θ_i^0 and the probability of other players' states is given by $b_i^0(\theta_{-i})$. (More precisely, $\mu^0(\theta_i^0, \theta_{-i}) = b_i^0(\theta_{-i})$ and $\mu^0(\theta) = 0$ if $\theta_i \neq \theta_i^0$.) Let μ^t be the probability distribution of $\theta(t)$, given μ^0. Then we have $\lim_{t \to \infty} \mu^t = \mu$ (μ is the invariant distribution with support X). Since μ assigns a strictly positive probability for any $\theta \in X$, so does μ^t for all sufficiently large t. Hence, for any $\theta \in X$ and any i, the conditional distributions given θ_i are well defined for all large t and we have

$$\lim_{t \to \infty} \mu^t(\theta_{-i}|\theta_i) = \mu(\theta_{-i}|\theta_i) \text{ for all } \theta \in X.$$

_____

is one.

Fix any $\theta \in X$ and let $H_i^t(\theta_i)$ be the set of possible private histories leading to state θ_i at time t: Formally, $H_i^t(\theta_i)$ is the set of private histories h_i^t such that (i) $\theta_i(0) = \theta_i^0$, $\theta_i(t) = \theta_i$, and (ii) h_i^t realizes with a positive probability under Markov chain $MC(\Theta, m)$ with initial distribution μ^0. The above argument shows that $H_i^t(\theta_i) \neq \varnothing$, for all sufficiently large t. Then, the conditional probability is well defined for all large t and expressed as

$$\mu^t(\theta_{-i}|\theta_i) = \frac{\Pr(\theta(t) = \theta)}{\Pr(\theta_i(t) = \theta_i)}$$

$$= \frac{\sum_{h_i^t \in H_i^t(\theta_i)} b_i(\theta_{-i}|h_i^t)\Pr(h_i^t)}{\sum_{h_i^t \in H_i^t(\theta_i)} \Pr(h_i^t),} \tag{6}$$

where $b_i(\theta_{-i}|h_i^t)$ is the conditional probability of $\theta_{-i}(t) = \theta_{-i}$ given private history h_i^t (this is equal to player i's belief after h_i^t).

Since automaton (m_i, θ_i^0) is optimal for player i at $t = 0$, following it after any private history which realizes with a positive probability should also be optimal. In particular, after private history $h_i^t \in H_i^t(\theta_i)$ player i is in state θ_i and therefore following automaton (m_i, θ_i) must be optimal after h_i^t. Note also that player i has belief $b_i(\cdot|h_i^t)$ after h_i^t. Those facts, taken together, imply that automaton (m_i, θ_i) is optimal under belief $b_i(\cdot|h_i^t)$, for any $h_i^t \in H_i^t(\theta_i)$.

Now let $B_i^{\theta_i}$ be the set of beliefs over θ_{-i} under which automaton (m_i, θ_i) is optimal for player i. Since (i) $B_i^{\theta_i} = \{b_i|W^*(b_i) = V_i^{\theta_i}(b_i)\}$ (ii) the optimal value function W^* is a continuous convex function and (iii) $V_i^{\theta_i}(\cdot)$ is a linear function, $B_i^{\theta_i}$ is a closed convex set (this may be stated as a lemma elsewhere. Also note that a direct proof, based on the linearity of expected payoff with respect to beliefs, is easy.). Equation (6) and our argument in the previous paragraph show that $\mu^t(\cdot|\theta_i)$ is a convex combination of the beliefs (i.e., $b_i(\theta_{-i}|h_i^t)$, $h_i^t \in H_i^t(\theta_i)$) in $B_i^{\theta_i}$. By the convexity of $B_i^{\theta_i}$, we have

$$\mu^t(\cdot|\theta_i) \in B_i^{\theta_i} \text{ for all } t \geq \bar{t}.$$

By the closedness of $B_i^{\theta_i}$ and $\lim_{t \to \infty} \mu^t(\cdot|\theta_i) = \mu(\cdot|\theta_i)$ implies

$$\mu(\cdot|\theta_i) \in B_i^{\theta_i} \text{ for all } \theta_i \in X_i.$$

This means that the joint distribution μ over initial state profile $\theta \in X$ induces a correlated equilibrium. ∎

6 Examples

6.1 Example 1

We apply the POMDP technique to the prisoner's dilemma model analyzed by Sekiguchi (1997) and Bhaskar and Obara (2002). The stage payoff is given by

	C	D
C	$1, 1$	$1 + g, -l$
D	$-l, 1 + g$	$0, 0$

Each player's private signal is $\omega_i = c, d$, which is a noisy observation of the opponent's action. For example, when the opponent chose C, player i is more likely to receive the correct signal $\omega_i = c$, but sometimes an observation error provides a wrong signal $\omega_i = d$. More precisely, we assume that exactly one player receives a wrong signal is $\varepsilon > 0$ and both receive wrong signals with probability $\xi > 0$. For example, when action profile is (C, C), the joint distribution of private signals is

	c	d
c	$1 - 2\varepsilon - \xi$	ε
d	ε	ξ

Hence in this model we have five parameters g, l, ε, ξ, and δ.

In Example 1, we assume that $\varepsilon = (1 - r)r$, $\xi = r^2$ (independent observation errors), $r = 1/6$, $g = 1$, $l = 1$, and $\delta = 0.9$. The candidate path preautomaton m_i corresponds to the grim trigger strategy and is shown in the following figure. We show that this preautomaton constitutes a correlated equilibrium, given some initial distribution of states. First, we determine $v^{\theta_i \theta_j}$, player i's payoff under state profile (θ_i, θ_j). This can be done by finding a solution to the system of equations

$$\begin{cases} v^{RR} = 1 + \delta \left\{ (1 - 2\varepsilon - \xi) v^{RR} + \varepsilon v^{RP} + \varepsilon v^{PR} + \xi v^{PP} \right\} \\ v^{RP} = -l + \delta \left\{ (\varepsilon + \xi) v^{RP} + (1 - \varepsilon - \xi) v^{PP} \right\} \\ v^{PR} = (1 + g) + \delta \left\{ (\varepsilon + \xi) v^{PR} + (1 - \varepsilon - \xi) v^{PP} \right\} \\ v^{PP} = 0 + \delta v^{PP} \end{cases} . \tag{7}$$

The solution shows

$$v^{RR} = 3.058\,8,\, v^{RP} = -1.176\,5,\, v^{PR} = 2.352\,9,\, v^{PP} = 0.$$

Let b denote player i's belief that j is in state R, and let $V^{\theta_i}(b)$ the expected payoff to player i when his state is $\theta_i = R, P$.[8] Then, the graphs of $V^{\theta_i}(b) = bv^{\theta_i R} + (1 - b)v^{\theta_i P}$, $\theta_i = R, P$ are given in the following figure. The horizontal axis measures b.

---

[8] In our general notation, $v^{\theta_i \theta_j} = v_i^{(m_i, \theta_i)\theta_j}$ and $V^{\theta_i} = V_i^{(m_i, \theta_i)}$.
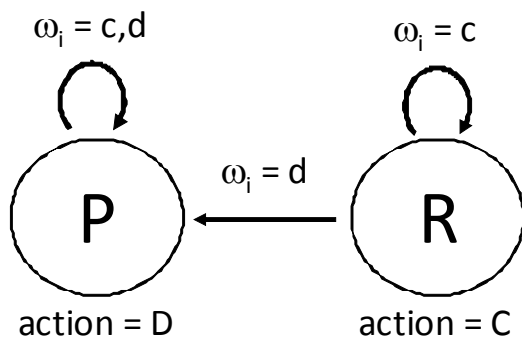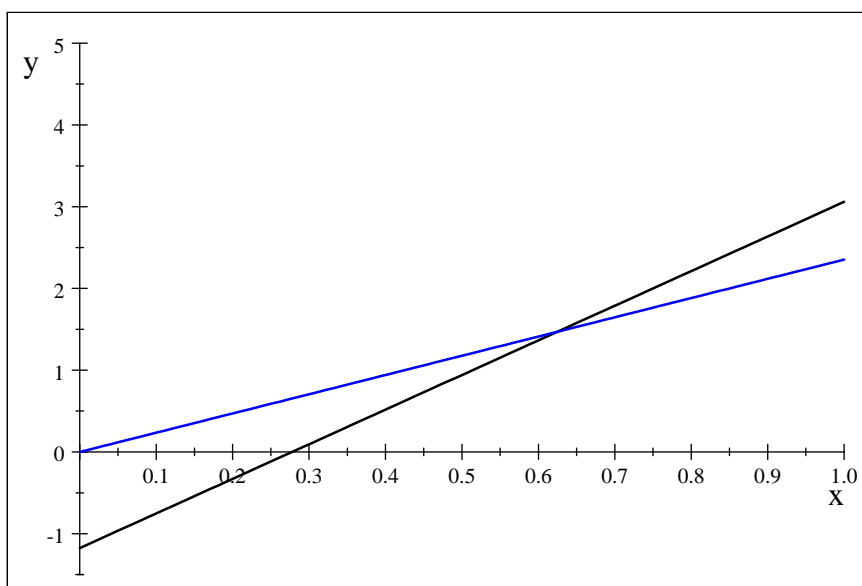
From top to bottom (on the y axis) V^P (blue) and V^R (black).

The initial value function in the POMDP iteration W^0 is the upper envelope of those two linear functions ($W^0(b) = \max\{V^R(b), V^P(b)\}$).

In the first step of POMDP, we consider one-shot extensions of path automata $\{(m_i, R), (m_i, P)\}$. Let $M^{a_i z z'}$ ($a_i = C, D$ and $z, z' = P, R$) be the one-shot ex-

tension that (i) starts with a state with action a_i, (ii) moves to state z of m_i after $\omega_i = c$, and (iii) moves to state z' of m_i after $\omega_i = d$. For example, M^{CRP} starts with action C and goes to state R (to play the grim trigger strategy) after $\omega_i = c$ and moves on to P (to play permanent defection) after $\omega_i = d$. It is easy to see that M^{CRP} actually implements the same strategy as (m_i, R) (the grim trigger strategy).

To perform the cream skimming of those one-shot extensions, looking at the belief dynamics is useful. Let $\chi_{a_i\omega_i}(b)$ denote the posterior probability of $\theta_j = R$ when the current action, signal, belief of player i are a_i, ω_i, and b. By Bayes' rule we have

$$\chi_{Cc}(b) = \frac{b(1 - 2\varepsilon - \xi)}{b(1 - \varepsilon - \xi) + (1 - b)(\varepsilon + \xi)},$$

$$\chi_{Cd}(b) = \frac{b\varepsilon}{b(\varepsilon + \xi) + (1 - b)(1 - \varepsilon - \xi)},$$

$$\chi_{Dc}(b) = \frac{b\varepsilon}{b(1 - \varepsilon - \xi) + (1 - b)(\varepsilon + \xi)}, \text{ and}$$

$$\chi_{Dd}(b) = \frac{b\xi}{b(\varepsilon + \xi) + (1 - b)(1 - \varepsilon - \xi)}.$$

For our parameter values in this example, it can be checked that $\chi_{ad}(b) < \chi_{ac}(b)$ for all b and $a = C, D$ (bad signal d always reduces the probability that the opponent is still cooperating). Let b^* be the point where V^R and V^P intersect. Since

$$W^0(b) = \begin{cases} V^R(b) & \text{if } b \geq b^* \\ V^P(b) & \text{if } b \leq b^* \end{cases},$$

$W^0(\chi_{ac}(b)) = V^P(\chi_{ac}(b))$ implies $W^0(\chi_{ad}(b)) = V^P(\chi_{ad}(b))$. This means that, for any b, the maximum value to achieve $\Gamma W^0(b)$ in the Bellman equation (2) is *not* achieved by M^{CPR} or M^{DPR}. Hence, for any b, $\Gamma W^0(b)$ is the associated with one of the following six automata:

$$M^{CRR}, M^{CRP}, M^{CPP}, M^{DRR}, M^{DRP}, M^{DPP}.$$

Note that M^{CRP} and M^{DPP} implement the same strategies as M^R (the grim trigger strategy) and M^P (permanent defection).

Let $V^{a_izz'}(b)$ be the expected payoff to player i, when he plays this automaton $M^{a_izz'}$ under belief b. This is a linear function in belief b:

$$V^{a_izz'}(b) = bv^{a_izz',R} + (1 - b)v^{a_izz',P}, \tag{8}$$

where $v^{a_izz',R}$ is the payoff under automaton $M^{a_izz'}$ when the opponent plays (m_j, R) ($v^{a_izz',P}$ is defined similarly).

To compute v^{a_izz',θ_j}, we can just use $V^R(b)$, $V^P(b)$ and belief functions $\chi_{a_i\omega_i}(b)$. For example, $v^{Czz',R}$ ($zz' = RR, PP$) is determined by
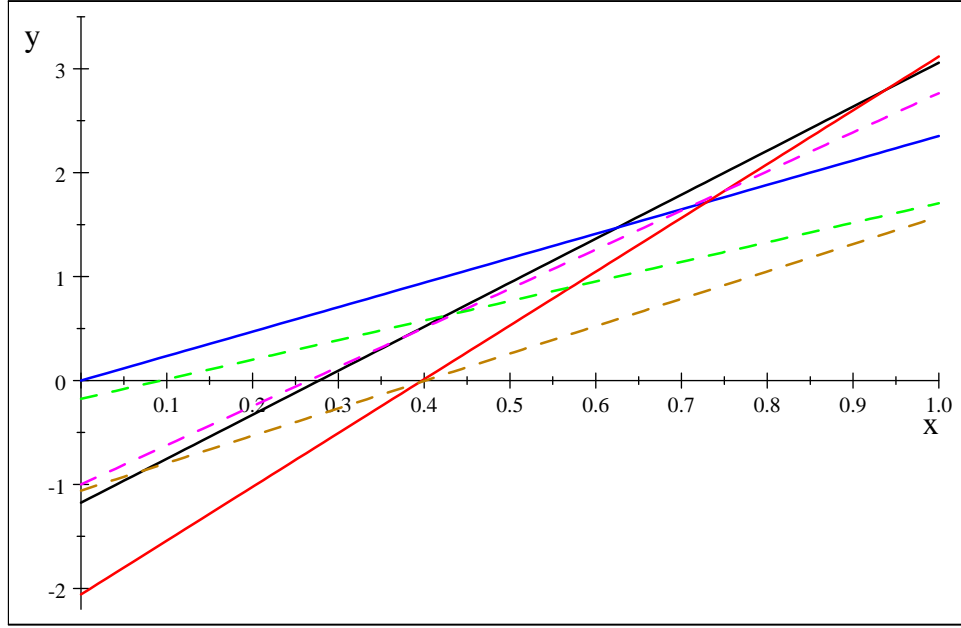
$$v^{Czz',R} = g_1(C, C) + \delta \left(V^z(\chi_{Cc}(1)) \Pr(\omega_i = c|CC) + V^{z'}(\chi_{Cd}(1)) \Pr(\omega_i = d|CC) \right).$$

Computation shows

$$v^{CRR,R} = 3.\,117\,6, v^{CPP,R} = 2.\,764\,7, v^{CRR,P} = -2.\,058\,9, v^{CPP,P} = -1,$$

$$v^{DRR,R} = 1.\,576\,4, v^{DRP,R} = 1.7059, v^{DRR,P} = -1.\,058\,9, v^{DRP,P} = -0.176\,48.$$

Then, the new value function $W^1(b)$ is the upper envelope of six linear functions $V^{a_i zz'}(b)$, $a_i zz' = CRR, CRP, CPP, DRR, DRP, DPP$. It is given by the following figure.

From top to bottom (on the y axis) $V^P = V^{DPP}$ (blue), V^{DRP} (green dotted), V^{CPP} (pink dotted), V^{DRR} (brown dotted), $V^R = V^{CRP}$ (black), and V^{CRR} (red).

POMDP to compute W^1

The upper envelope W^1 is almost the same as the original value function W^0 (= the upper envelope of the blue and black lines). Note that the red line (V^{CRR}) coincides with W^1 only for very high belief $b \in [0.937\,53, 1]$, and one can check that $\chi_{a_i \omega_i}(b)$ does not lie in this region for any a_i, ω_i and b. This implies that $W^1(\chi_{a_i \omega_i}(b)) = W^0(\chi_{a_i \omega_i}(b))$ for all a_i, ω_i, and b, which shows $\Gamma W^1 = \Gamma W^0 = W^1$. Hence POMDP converged to the optimal value function $W^* = W^1$ on $\Theta = [0, 1]$ in two steps.

Note that we can find a fixed point one step earlier if we restrict attention to a belief-closed set $X_i = [0, \frac{14}{29} \, (\approx 0.79167)]$, where $\frac{14}{29}$ is a fixed point of χ_{Cc} in $(0, 1)$.

If the current belief is in X_i, then every posterior belief is in X_i given any (a_i, ω_i) because there is always a positive probability that a bad signal is observed even when C is played. Since V^{CRR} matters only for $b_i \geqq 0.93753$, W^0 is the fixed point on X_i.

**Conclusion:** The above figure and our computation show

1. the optimal value function W^* coincides with V^P for $b \in [0, 0.625]$

2. the optimal value function W^* coincides with V^R for $b \in [0.625, 0.937\,53]$

Hence, when a correlation device can generate initial beliefs $b^R \in [0.625, 0.937\,53]$ and $b^P \in [0, 0.625]$, we get a correlated equilibrium where players use automata M^R (the grim trigger strategy) and M^p (permanent defection). One such equilibrium is the mixed strategy equilibrium identified by Bhaskar and Obara, in which the initial belief is given by 0.625 (where the player is indifferent between M^R and M^P).

6.2 Example 1 - A Simpler Verification

The previous section illustrates how POMDP works, but for this particular example, there is a much simpler verification of equilibrium conditions, based on Proposition 5 (*Finite Verification of the Belief-Based One-Shot Deviation Principle*). Let $b^c = \frac{14}{29}$ (≈ 0.79167) be the fixed point of χ_{Cc}. Then, one can see that $X = [0, b^c]$ is a belief-closed set. The reason is the following. If we start with $b \in X \backslash \{0\}$ and play C and observe $\omega_i = c$ repeatedly, the posterior belief increases and approaches the fixed point b^c. This is because the graph of posterior belief function χ_{Cc} is increasing and cross the 45° degree line at b^c from above (i.e., b^c is a stable fixed point). If current action and signal is not (C, c), the posterior belief goes down. Finally, $b = 0$ is absorbing. Hence X is belief-closed.

We will show that $\Gamma W^0 = W^0 = \max_{\theta_i = R, P} V^{\theta_i}$ on X. Proposition 5 shows that we only need check this at three points, $0, b^*$(the belief at which V^P and V^R cross), and b^c.

One needs to check $W^0 = \Gamma W^0$ on only three points 0, b*, and b^c.
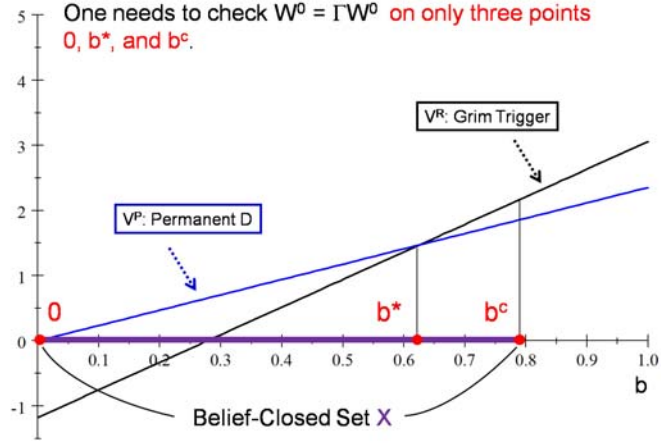
Checking $\Gamma W^0(0) = W^0(0)$ is trivial, because $b = 0$ means that the opponent is playing D forever and the best reply is permanent defection. To check $\Gamma W^0(b^*) = W^0(b^*)$, we check condition (1) in Proposition 5. If D is played today, the posterior belief is always less than b^* and the optimal continuation (according to W^0) is permanent defection. This is exactly what automaton (m_i, P) specifies. Now suppose C is played today. If the private signal is c, the posterior belief moves up (it is above b^*), and the optimal continuation (according to W^0) is grim trigger. If the private signal is d, the posterior belief goes down (it is below b^*), and the optimal continuation (according to W^0) is permanent defection. Again, this is exactly what automaton (m_i, R) specifies. Hence condition (1) of Proposition 5 is satisfied.

Lastly, let us check $\Gamma W^0(b^c) = W^0(b^c)$. We know $\chi_{Cc}(b^c) = b^c$ and calculation shows $\chi_{Cd}(b^c) < b^*$, so condition (1) is satisfied. When D is played today, $\chi_{Dc}(b^c)$ and $\chi_{Dd}(b^c)$ are both below b^*, and therefore the maximum value of one-shot extension which plays D is given by $V^P(b^c)$ (permanent defection: the blue line in the figure). Since this is less than $V^R(b^c)$, the condition (2) of Proposition 5 is satisfied.

Hence, the analysis in Sekiguchi (1997) and Bhaskar and Obara (2002) can be much simplified as above, according to our general belief-based methodology.

6.3 Example 2

Modify the stage game of the previous example as follows:

	C	$C\prime$	D
C	$1, 1$	$1, 1 - \varepsilon$	$-l, 1 + g$
C'	$1 - \varepsilon, 1$	$1 - \varepsilon, 1 - \varepsilon$	$-l - \varepsilon, 1 + g$
D	$1 + g, -l$	$1 + g, -l - \varepsilon$	$0, 0$

You can interpret C' as C plus some monitoring activity that costs $\varepsilon > 0$. If a player chooses C', then he can observe the other player's action perfectly. (This violates

our full support assumption, but the analysis below remains essentially the same even when we assume that a player with action C' can observe the other player's action *almost* perfectly.) From the other player, C and C' are indistinguishable, i.e. the other player observes c with probability $1-r$ and d with probability r whether C or C' is played. We show that the grim trigger strategy still constitutes a correlated sequential equilibrium for some initial distribution.

Let's use the same parameter values for g, l, r and δ as before. Let V^R and V^P be the value function associated with R and P as before. In the first step of POMDP, we need to consider eight one-shot extensions this time: $M^{a_i, z_i, z_i'}, a_i \in \{C, C', D\}, z_i, z_i' \in \{R, P\}$. Again we don't have to consider a combination of $(z_i, z_i') = (R, P)$. Let $V^{a_i, z_i, z_i'}$ be the discounted payoff function associated with $M^{a_i, z_i, z_i'}$. We already know that $V^{CPP}, V^{DRR}, V^{DRP}$ are dominated. We have three new payoff functions because of $C' : V^{C'RR}, V^{C'RP}, V^{C'PP}$. It is easy to show that we only need to consider $V^{C'RP}$ (remember that monitoring is perfect given C').

How does this new function $V^{C'RP}$ affect the updated value function? Suppose that $\varepsilon = 0$ for the time being. Then $M^{C'RP}$ dominates M^{CRR} and M^{CRP} because C' generates the same expected payoff as C in the current period, but more informative than C. So $V^{C'RP}$ is above V^{CRR} or V^{CRP}. Only at $b = 1$, $V^{C'RP}$ coincides with V^{CRR} because the player's signal does not convey any information about the the other player's current state (which he knows to be R). Thus $V^{C'RP}$ looks as follows.[9]
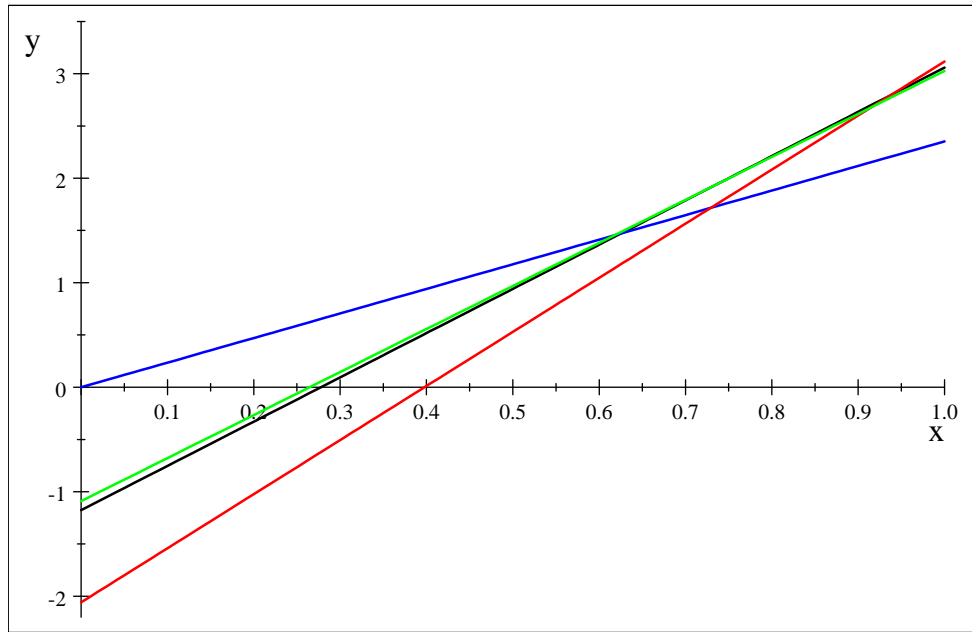
---

[9]Note that $v^{C'RP,R} = v^{CRP,R}$ and $v^{C'RP,P} = -l = -1$. Hence (with $\varepsilon = 0$),

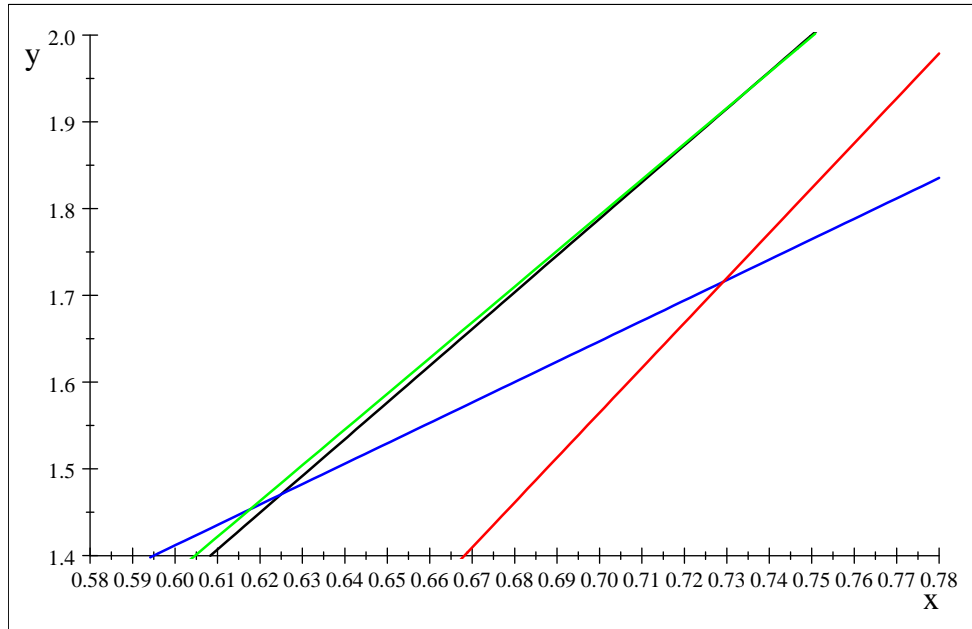$$V^{C'RP}(b) = bv^{C'RP,R} + (1-b)v^{C'RP,P} = 3.1176b - (1-b).$$
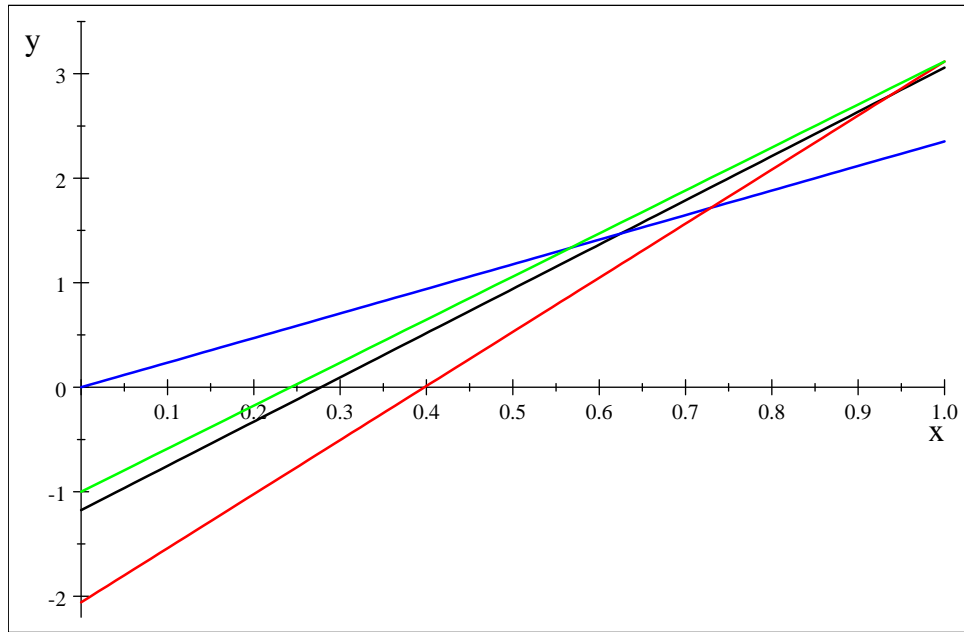
From top to bottom (on the y axis) $V^P = V^{DPP}$ (blue), $V^R = V^{CRP}$ (black), V^{CRR} (red), and $V^{C'RP}$(green) (with $\varepsilon = 0$).

Now let's set $\varepsilon = 0.09$. Then the green line just shifts down by ε, but still above the blue line and the green line around 0.6 as follows.

From top to bottom (on the y axis) $V^P = V^{DPP}$ (blue), $V^R = V^{CRP}$ (black), V^{CRR} (red), and $V^{C'RP}$ (green) (with $\varepsilon = 0.09$).

If we magnify the range around 0.625 (the intersection of the black and the blue line), we have the following graph.

From top to bottom (on the y axis) $V^P = V^{DPP}$ (blue), $V^R = V^{CRP}$ (black), V^{CRR} (red), and $V^{C'RP}$ (with $\varepsilon = 0.09$).

The green line intersects with the black line (V^{CRP}) at $b = 0.73492$ and intersects with the blue line (V^{DPP}) at $b = 0.61767$.

This means that $M^{C'RP}$ is better than M^{CRP} or M^{DPP} in some range of beliefs. Intuitively, when a player is near indifferent between playing the standard trigger strategy and "always D" strategy, he has incentive to pay ε to get a better information to make a more informed choice between M^{CRP} and M^{DPP}.

Since d is observed with probability $\frac{1}{6}$ by the other player even when a player plays C or C', posterior beliefs never exceeds $\frac{5}{6}$. So $X_i = [0, \frac{5}{6}]$ is a belief-closed set. We show that a fixed point can be found in two steps on X_i.

In the first step of POMDP, we obtain the value function $W^1 = \Gamma W^0$ as an upper envelope of the four linear functions: $V^R = V^{CRP}, V^P = V^{DPP}, V^{CRR}$ and $V^{C'RP}$. The value function is pushed up a bit around $b_i = 0.625$ by $V^{C'RP}$. In the second step, $W^2 = \Gamma W^1$ is an upper envelope of all linear functions that survived in the previous step as well as new linear functions such as $V^{a_i RM}$, where M means $M^{C'RP}$ is played upon an observation of d. But it can be shown that every such new linear function is dominated by one of old linear functions. This is based on the following observations. Recall V^R crosses V^P at $b = 0.625$. First, the posterior belief may never fall in $[0.61767, 0.73492]$ when M^{CRP} is played in $[0.625, \frac{5}{6}]$ or M^{DPP} is played in $[0, 0.625]$ ($\chi_{Cc}(0.625) = 0.74405, \chi_{Cd}\left(\frac{5}{6}\right) = 0.41667$).[10] Second, the value function in $[0.61767, 0.73492]$ matters only when C is played for some range of beliefs $[0, 0.625]$, but C must be still suboptimal action for such beliefs because the continuation value is improved only slightly. Hence the value function obtained in the second step is the same one as the one obtained in the first step because it is the upper envelope of the same set of linear functions. Therefore we have a fixed point $W^2 = \Gamma W^1 = W^1$.

6.4 Example 3:

Consider a repeated prisoners' dilemma with imperfect public monitoring. The stage payoff is given by

	C	D
C	$1, 1$	$-0.6, 1.2$
D	$1.2, -0.6$	$0, 0$

.

---

[10] However, the posterior belief after Cd falls in $[0.6245, 0.6371]$ when the prior belief is very high. For such prior beliefs, the optimal plan is M^{CRM} rather than M^{CRR}. This means that the value function is moving upward strictly for high beliefs. It is not difficult to see that we don't find a quick convergence in this area. However, this area is outside of the belief-closed set we focus on. This is why we focus on the belief-closed set $X_i = [0, \frac{5}{6}]$.

where V^{CRR} was increasing in

Let y be a public signal, which takes the value of either G or B. Assume that the monitoring structure is as follows:

	G	B
CC	$\frac{3}{4}$	$\frac{1}{4}$
CD or DC	$\frac{1}{2}$	$\frac{1}{2}$
DD	$\frac{1}{4}$	$\frac{3}{4}$

For example, the first row means that G realizes with probability $\frac{3}{4}$ when (C,C) is taken.

We turn this game into a game with private monitoring by introducing private observation noise. We assume that player i observes a private signal $s_i \in \{g,b\}$ instead of y. When both players take the same action, their signals are perfectly correlated with y, that is, they observe the true public signal with probability 1 i.e. $\Pr(s_i = g|y = G, CC \text{ or } DD) = \Pr(s_i = b|y = B, CC \text{ or } DD) = 1$ for $i = 1, 2$. However, when they take different actions such as (C,D) or (D,C), their private signals are uncorrelated with y, and g and b are observed with equal probability independently by each player independent of realizations of y. More precisely, the distributions of their private signals are given by $\Pr(s_i = g|y = G, CD \text{ or } DC) = \Pr(s_i = b|y = B, CD \text{ or } DC) = \frac{1}{2}$ (more generally we can set this equal to $1 - \varepsilon \in \left[\frac{1}{2}, 1\right)$ to allow imperfect correlation between y and s_i given (C,D) or (D,C)).

Tit-for-Tat is a partial strategy in which a player plays what he observed in the previous period. This strategy can be represented by the following two-state preautomaton

$$
\begin{aligned}
\Theta_i &= \{R, P\} \\
f_i(R) &= C, f_i(P) = D \\
T_i(R|R, g) &= T_i(R|P, g) = T_i(P|R, b) = T_i(P|P, b) = 1.
\end{aligned}
$$

We use POMDP to show that Tit-fot-Tat can be a finite state equilibrium.

Let $V_i^{zz'}$ be player i's discounted payoff when he is in state z and the other player is in state z'. These values can be derived by solving the following system of equations:

$$
\begin{aligned}
V^{RR} &= 1 + \delta\left[\frac{3}{4}V^{RR} + \frac{1}{4}V^{PP}\right] \\
V^{RP} &= -0.6 + \delta\left[\frac{1}{4}V^{RR} + \frac{1}{4}V^{RP} + \frac{1}{4}V^{PR} + \frac{1}{4}V^{PP}\right] \\
V^{PR} &= 1.2 + \delta\left[\frac{1}{4}V^{RR} + \frac{1}{4}V^{RP} + \frac{1}{4}V^{PR} + \frac{1}{4}V^{PP}\right] \\
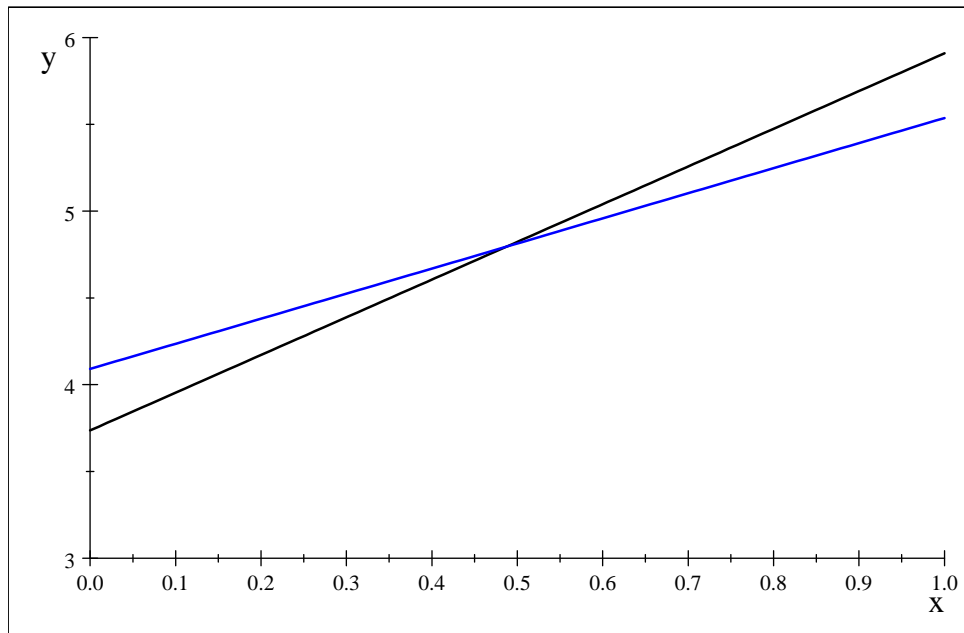V^{PP} &= \delta\left[\frac{1}{4}V^{RR} + \frac{3}{4}V^{PP}\right]
\end{aligned}
$$

So

$$V^{RR} = \frac{4 - 3\delta}{2(1-\delta)(2-\delta)}$$

$$V^{RP} = -0.6 + \delta\left(\frac{1}{4(1-\delta)} + \frac{1}{4}\frac{2(1-\delta)(1.2-0.6)+\delta}{(1-\delta)(2-\delta)}\right)$$

$$V^{PR} = 1.2 + \delta\left(\frac{1}{4(1-\delta)} + \frac{1}{4}\frac{2(1-\delta)(1.2-0.6)+\delta}{(1-\delta)(2-\delta)}\right)$$

$$V^{PP} = \frac{\delta}{2(1-\delta)(2-\delta)}$$

Assume $\delta = 0.9$. Then

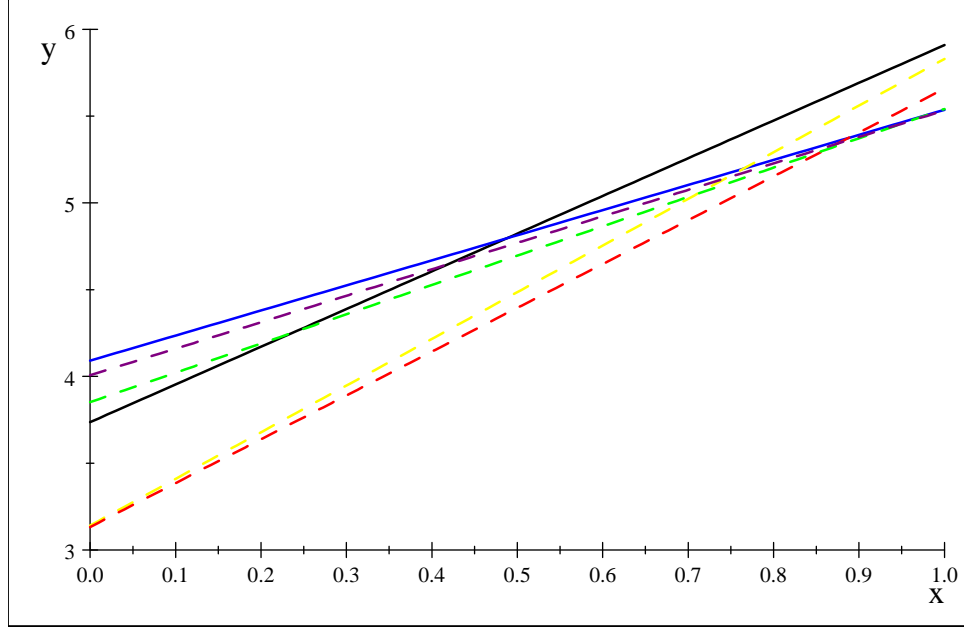$$V^{RR} = 5.9091, \ V^{RP} = 3.7364, \ V^{PR} = 5.5364, \ V^{PP} = 4.0909.$$

Let b_i be player i's belief that player j is in state R and $V_i^Z(b_i) = b_i V^{ZR} + (1 - b_i)V^{ZP}$. V_i^R and V_i^P are plotted below.

V^R (black), V^P (blue).

These two lines intersect at $b^* = 0.48749$. W^0 is the upper envelope of these two functions. $W^1 = \Gamma W^0$ can be obtained by consider M. In the first step of POMDP,

we compute $V^{a_i z_i z_i}$ easily as before.[11] They are plotted below.

$V^R = V^{CRP}$ (black), $V^P = V^{DRP}$ (blue), V^{CRR} (yellow), V^{CPP} (red), V^{DRR}(green), V^{DPP}(purple).

Since the black line and the blue line still constitute the upper envelope of these six linear functions, we have $\Gamma W^0 = W^0$ everywhere in $\Theta = [0, 1]$ in one step. Clearly these TFT path automata constitute a sequential equilibrium when the initial state is given by (R, R).

---

[11] The formulas are here for the record.

$$
\begin{aligned}
v^{CRR,R} &= 1 + 0.9(\frac{3}{4}V^{RR} + \frac{1}{4}V^{RP}) = 1 + 0.9(\frac{3}{4}5.9091 + \frac{1}{4}3.7364) = 5.8293 \\
v^{CRR,P} &= -0.6 + 0.9(\frac{1}{2}V^{RR} + \frac{1}{2}V^{RP}) = -1.2 + 0.9(\frac{1}{2}5.9091 + \frac{1}{2}3.7364) = 3.1405 \\
v^{CPP,R} &= 1 + 0.9(\frac{3}{4}V^{PR} + \frac{1}{4}V^{PP}) = 1 + 0.9(\frac{3}{4}5.5364 + \frac{1}{4}4.0909) = 5.6575 \\
v^{CPP,P} &= -0.6 + 0.9(\frac{1}{2}V^{PR} + \frac{1}{2}V^{PP}) = -1.2 + 0.9(\frac{1}{2}5.5364 + \frac{1}{2}4.0909) = 3.1323 \\
v^{DRR,R} &= 1.2 + 0.9(\frac{1}{2}V^{RR} + \frac{1}{2}V^{RP}) = 1.2 + 0.9(\frac{1}{2}5.9091 + \frac{1}{2}3.7364) = 5.5405 \\
v^{DRR,P} &= 0.9(\frac{3}{4}V^{RP} + \frac{1}{4}V^{RR}) = 0.9(\frac{3}{4}3.7364 + \frac{1}{4}5.9091) = 3.8516 \\
v^{DPP,R} &= 1.2 + 0.9(\frac{1}{2}V^{PR} + \frac{1}{2}V^{PP}) = 1.2 + 0.9(\frac{1}{2}5.5364 + \frac{1}{2}4.0909) = 5.5323 \\
v^{DPP,P} &= 0.9(\frac{3}{4}V^{PP} + \frac{1}{4}V^{PR}) = 0.9(\frac{3}{4}4.0909 + \frac{1}{4}5.5364) = 4.007
\end{aligned}
$$

Note: What would happen with Phelan and Skrzypacz approach in this example?

There are two approaches in Phelan and Skrzypacz (2009). The first approach is to compute $\overline{M}(\omega)$ recursively and check one shot deviation incentive constraints at boundary points of $\overline{M}(\omega)$ for each state ω. The second approach is to find $M^I(\omega)$ for each state ω and apply T^I to M^I repeatedly until (hopefully) it converges to some nonempty set. It turns out that both approaches fail in this example because the extended automaton version of TFT cannot be an equilibrium for any initial joint distribution.

- 1st approach:

  $\overline{M}(R) = [\frac{1}{2}, 1]$, $\overline{M}(P) = [0, \frac{1}{2}]$ are obtained by one iteration (to be precise, it is confirmed to be a fixed point in the second iteration). Remember that $b^* = 0.48749$, hence M^{DRP} is not optimal at $b = \frac{1}{2}$. Thus one-shot deviation constraint is clearly violated for $\overline{M}(P)$ at $\mu = \frac{1}{2}$.

- 2nd approach:

  $M^I(\omega)$ is the set of beliefs where one-shot deviation constraint at state ω is satisfied. With the full TFT, one-shot deviation from starting at R or P is the same as starting at P or R respectively. Hence $M^I(R) = [b^*, 1]$ and $M^I(P) = [0, b^*]$ by definition of b^*. Next the operator T^I is applied to M^I. Basically we eliminate any belief if it leads to inconsistent belief after some combination of action and signal in the next period. More specifically, we eliminate μ if either $\chi[a, g, \mu] \notin M^I(R)$ for any a or $\chi[a, b, \mu] \notin M^I(P)$ for any a. Note that, for μ close to 1, $\chi[D, b, \mu]$ is close to $\frac{1}{2}$, hence above b^* and not in $M^I(P)$. So such belief is eliminated. Let $\overline{\mu}$ be the infimum of such belief i.e. $\chi[D, b, \overline{\mu}] = \mu^*$. Then $(\overline{\mu}, 1]$ is eliminated. A similar problem arises when μ is close to 0. Let $\underline{\mu}$ be the belief such that $\chi[C, b, \underline{\mu}] = b^*$ (so $\chi[C, b, \mu] > b^*$ when $\mu \in [0, \underline{\mu}]$). Then $[0, \underline{\mu})$ is eliminated. It is straightforward to show that all other beliefs are kept in the first iteration. Hence $T^I(M^I)(R) = [\mu^*, \overline{\mu}]$ and $T^I(M^I)(P) = [\underline{\mu}, \mu^*]$. Observe that unraveling occurs from the second iteration on (note that $\chi[C, g, \mu]$ is increasing in μ, $\chi[C, g, 1] = 1$ and $\chi[C, g, \mu] > \mu$ for all μ). In fact, it is easy to see that $\lim_{n \to \infty} (T^I)^n(M^I)$ is an empty set (ϕ, ϕ).

6.5 Example 4

Modify Example 2 as follows. The stage payoff functions remain the same:

	C	C'	D
C	$1, 1$	$1, 1 - \varepsilon$	$-l, 1 + g$
C'	$1 - \varepsilon, 1$	$1 - \varepsilon, 1 - \varepsilon$	$-l - \varepsilon, 1 + g$
D	$1 + g, -l$	$-1 + g, l - \varepsilon$	$0, 0$

You can interpret C' as C plus some type of monitoring activity that costs $\varepsilon > 0$. If a player chooses C', then he can (1) observe the other player's action perfectly and (2) *make sure that the other player observes c with probability 1.* (This violates our full support assumption, but all of our theoretical results above are valid, if the full support assumption holds for the action profiles that are played on the equilibrium path. In Example 4, C' is not played on the equilibrium path.) Our goal is to show that the grim trigger strategy still constitutes a correlated sequential equilibrium for some initial distribution.

We use the same parameter values for g, l, r and δ as before and set $\varepsilon = 0.685$. Let $v^{zz'}$ be a player's discounted payoff when his state is z and his opponent's state is z'. Since C' is never played by the grim trigger strategy (m, R) and the perpetual defection (m, P), the following values are unchanged:

$$v^{RR} = 3.0588, v^{RP} = -1.1765, v^{PR} = 2.3529, v^{PP} = 0.$$

Let's apply POMDP.

1st step

In the first step, we compute the upper envelope of *nine* linear functions V^{CRR}, V^{CRP}, V^{CPP}, V^{DRR}, V^{DRP}, V^{DPP}, $V^{C'RR}$, $V^{C'RP}$, and $V^{C'PP}$ (again we don't have to consider the cases with $(z, z') = (P, R)$). The last three linear functions are new as C' is played in the first period. We already know that $V^{CPP}, V^{DRR}, V^{DRP}$ are dominated. Note that a player can infer the opponent's state perfectly when playing C' : the opponent's state is P if d is observed (by perfect monitoring) and the opponent's state is R if c is observed (by perfect monitoring & forcing the opponent to observe c). Hence the optimal continuation on-path automaton is (m, R) if c is observed and (m, P) if d is observed. So we only need to consider $V^{C'RP}$ among three new functions. The discounted value $V^{C'RP}$ associated with $M^{C'RP}$ is given by $V^{C'RP}(b) = b\left(1 - 0.685 + 0.9v^{CRP,R}\right) + (1 - b)(-1 - 0.685)$, which is given by the green line below
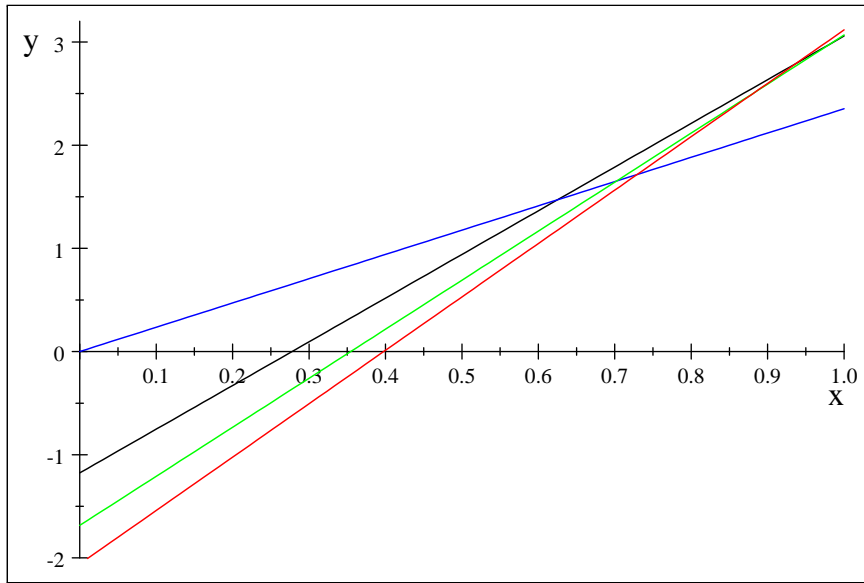
where V^{CRR} is the red line, V^{CRP} is the black line, and V^{CPP} is the black line as before.

It is not clear from the above graph, but the green line is not a part of the upper envelope. If we magnify the range around 0.9, we find the following picture.
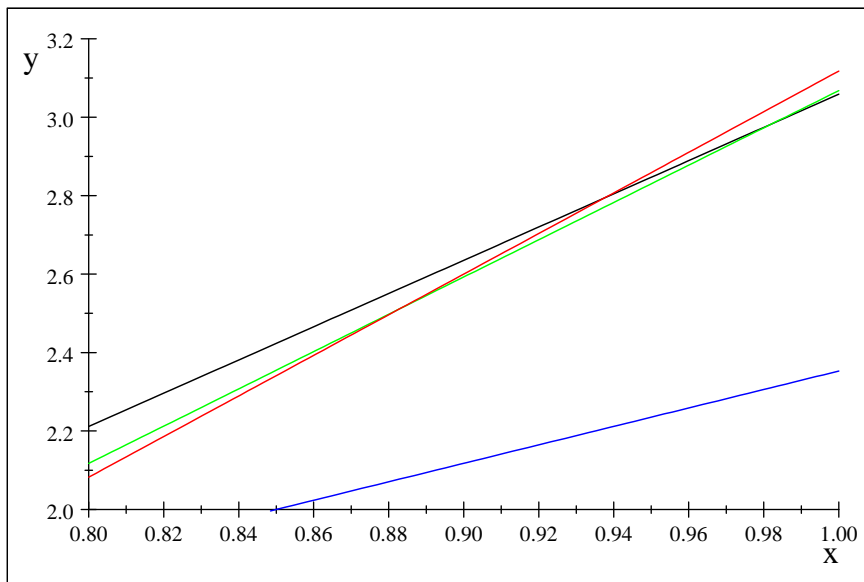
Hence we obtain exactly the same value function W^1 as in Example 1 in the first step of POMDP. We also have the same $\mathcal{M}^1 = \{M^{CRP}\,(= (m, R))\,, M^{DPP}\,(= (m, P))\,, M^{CRR}\}$ in the end of this step.

2nd step

Since W^1 is the same, the upper envelope we would obtain in the second step without C' is exactly the same as before: W^1 itself (remember that W^1 is the fixed point in Example 1 in the 2nd step). So the question is whether C' can affect this upper envelope. This time we have three on-path automatons to choose from for the continuation play: $M^{CRP}\,(= (m, R))\,, M^{DPP}\,(= (m, P))$ and M^{CRR}. Let's denote this last automaton by M_1. Since M_1 is optimal among these three automatons given $b = 1$, we only need to consider $M_2 = M^{C'M_1P}$, which is a one-step extension that starts with playing C' and moves to M_1 given c and P given d. The discounted value V^{M_1} associated with M_2 is given by $V^{M_2}(b) = b\left(1 - 0.685 + 0.9v^{CRR,R}\right) + (1-b)\,(-1 - 0.685)\,.$ Note that a player's continuation value after (C', c) is $v^{CRR,R}$, which is larger than $v^{CRP,R}$ in the first step. So the value of M_2 generates a larger payoff than $M^{C'RP}$. In fact, V^{M_2} (green line) dominates V^{CRR} as the following picture shows.
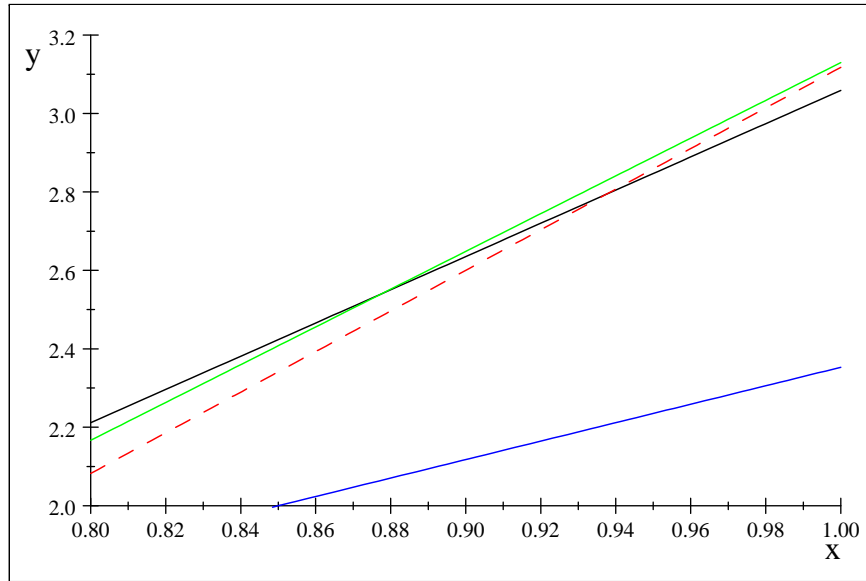
The green line intersects with the black line at $b = 0.87742$. Therefore W^2 is the upper envelope of V^{CRP}(black), V^{DPP}(blue) and V^{M_2}(green) and $\mathcal{M}^2 = \{M^{CRP}, M^{DPP}, M_2\}$ in this step. We don't get a convergence in the 2nd step yet.

3rd step

In the third step, the black line and the blue line do not change. This is because the posterior belief never falls in the area where the green line is a part of the upper envelope (the posterior belief is very low given D, and the posterior belief given C converges to the fixed point of χ_{Cc}, which is $\frac{14}{29}$ (≈ 0.79167)). However, the green line moves up slightly because $W^2(b) > W^1(b)$ at $b = 1$. The automaton generating this new green line is $M_3 = M^{C'M_2P}$, which is one-step extension that plays C' and moves to M_2 given c and P given d. Now the green line intersects with the black line at $b = 0.87743$. Other than that, we have exactly the same picture.
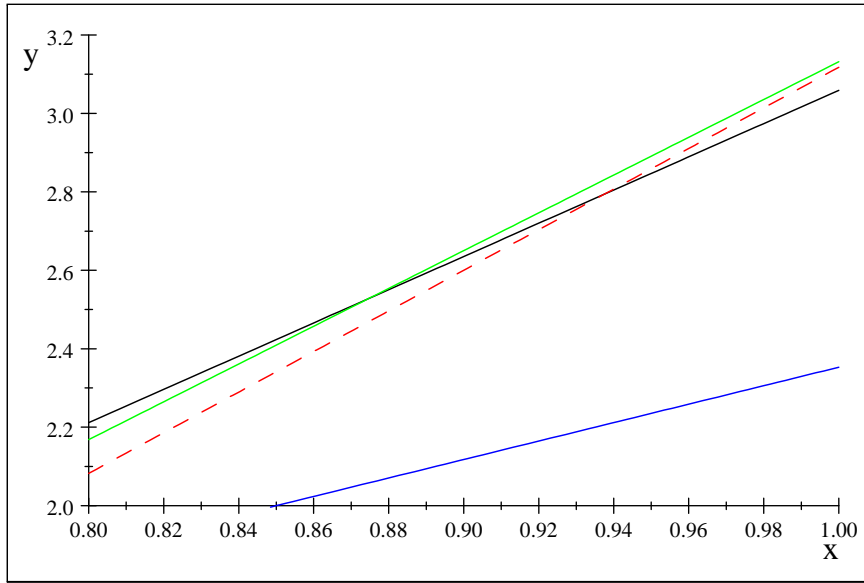
W^3 is the upper envelope of the same black and blue line, and a slightly higher green line. We have $\mathcal{M}^3 = \{M^{CRP}, M^{DPP}, M_3\}$

...30th step

After the third step, the relative locations of the linear functions remain the same: only the green line moves up slightly in each step, whereas the black line and the blue line stays the same. For example, we have the following picture in the 30th step.
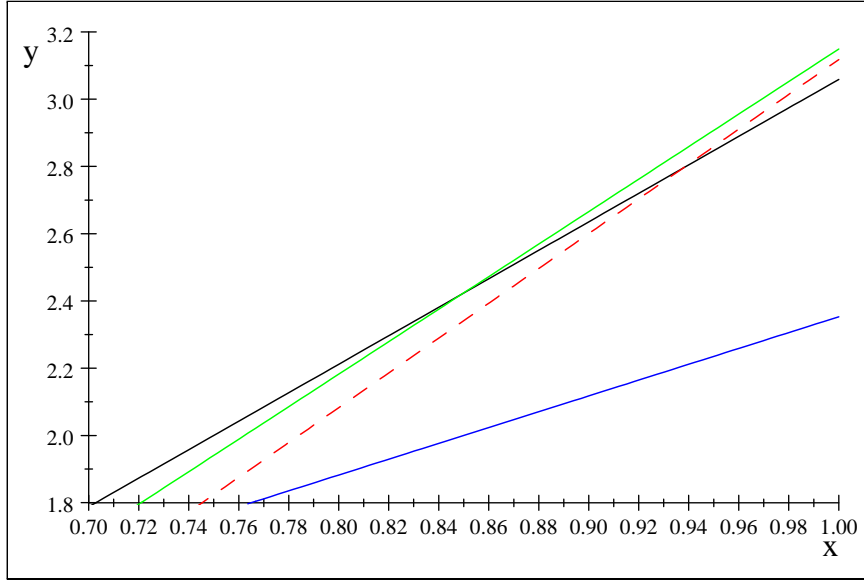
Now the green line intersects with the black line at $b = 0.84933$. The automaton generating the green line is M_{30} that plays C' 29 times before moving to M^{CRR} as long as d is not observed. The upper envelope W^{30} still consists of the black, blue and the green line and $\mathcal{M}^{30} = \{M^{CRP}, M^{DPP}, M_{30}\}$. We don't obtain a convergence yet and clearly this process still continues (in fact convergence is never obtained in a finite number of steps.)

So what should we do?

We apply our Proposition 4 here. Let $b^* = 0.625$ be the belief where V^P and V^R crosses and $\overline{b} = \frac{14}{29} = 0.79167$ be the fixed point of χ_{Cc}. It is easy to verify that the following collection of belief sets constitute an on-path belief closed set.

$$X_i(P) = [0, b^*], \ X_i(R) = [b^*, 0.79167].$$

We show that (4) is satisfied for any belief on this on-path closed set given $\mathcal{M}^{30}$ and W^{30} we obtained in the 30th step. . Deviating to D given $b_i \in X_i(R)$ or deviating to C given $b_i \in X_i(P)$ does not move the posterior belief outside of the on-path belief closed set. The fact that these deviations do not improve the payoff was already verified many times using the value function $W_0 = W_1 =, ..., = W_{29} \left(= W^{30} \right)$ on this on-path belief closed set. So we can focus on a deviation to C', which moves the posterior belief outside of the on-path belief closed set (to $b = 1$) when c is observed.

The only relevant one-shot extension of $\mathcal{M}^{30} = \{M^{CRP}, M^{DPP}, M_{30}\}$ that start with C' is M_{31}, which plays C' today followed by M_{30}. We know that the $W^{30}(1)$

was obtained in the 30th step, which is 3.1489.[12] The value of path automaton $M_{31} \in \widetilde{\mathcal{M}}^{30}$ is given by

$$V^{M_{31}} = b(1 - 0.685 + 0.9W^{30}(1)) + (1 - b)(-1 - 0.685).$$

The value adjusting component in Proposition 4 (4) is $p(M_{31}, X_i, b)\frac{\delta^{31}}{1-\delta}|W_1 - W_0| = b \times \delta \times \frac{\delta^{30}}{1-\delta}|W_1 - W_0| = b \times 0.9 \times \frac{0.9^{30}}{1-0.9}0.06 = b \times 0.9 \times 0.025435$, because $p(M_{31}, X_i, b) = b$. Recall that, when C' is chosen (by M_{31}) and the opponent is in state R (which happens with probability b), the opponent remains in R with probability one (hence the posterior belief $b' = 1$ is outside of the on-path belief-closed set X_i). Therefore,

$$
\begin{aligned}
& V^{M_{31}} + p(M_{31}, X_i, b)\frac{\delta^{31}}{1-\delta}|W_1 - W_0| \\
=\ & b\left(1 - 0.685 + 0.9\left(W^{30}(1) + 0.025435\right)\right) + (1 - b)\left(-1 - 0.685\right), \\
=\ & b\left(1 - 0.685 + 0.9\left(3.1743\right)\right) + (1 - b)\left(-1 - 0.685\right).
\end{aligned}
$$

This adjusted value function is depicted by the light green line in the following figure.

---
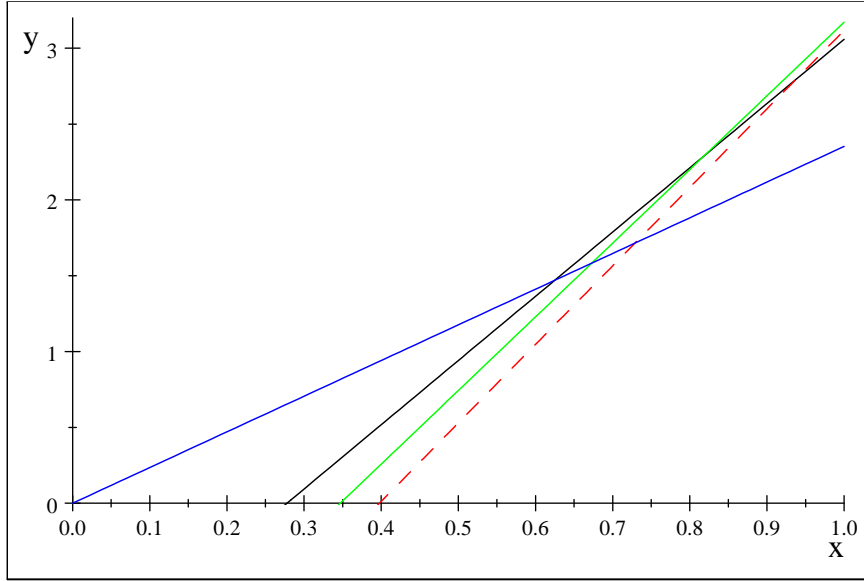
[12] The general formula is

$$
\begin{aligned}
W^n(1) &= \frac{1 - \delta^{n-1}}{1 - \delta}(1 - 0.685) + \delta^{n-1}W^1(1) \\
&= \frac{1 - 0.9^{n-1}}{0.1}(1 - 0.685) + 0.9^{n-1}v^{CRR,R}.
\end{aligned}
$$

Observe that the green line does not constitute a part of the upper envelope on $[0, 0.79167]$. Hence (4) is satisfied at every belief on this on-path belief-closed set. Therefore (m, R) is the optimal strategy for any $b_i \in X_i(R)$ and (m, P) is the optimal strategy for any $b_i \in X_i(P)$ by Proposition 4.

7 Discussion and Comparison to Phelan and Skrzypacz (2009)

Let $\mathcal{M}_i$ be the set of all path automata for player i, and define a **belief-based continuation path** $\beta_i : \Delta(\Theta_{-i}) \to \mathcal{M}_i$. This specifies the intended path of play for each initial belief (but it does not specify what to do after deviating from the intended path in the future). At first glance, this may not seem to be a natural or useful object, compared to a repeated game strategy $s_i : H_i \to A_i$ or a belief-based optimal policy (action) function $\gamma_i : \Delta(\Theta_{-i}) \to A_i$. A message of the present paper, however, is that a belief-based continuation path β_i is the right concept to focus on in the belief-based analysis, and our method verifies the optimality of β_i as follows. Let $\widetilde{\mathcal{M}}_i(\beta_i)$ be the set of one-shot extensions of the path automata specified by β_i (i.e., $\beta_i(\Delta(\Theta_{-i}))$). Our method presented in Section 5.1 may be rephrased as:

Belief-Based One-Shot Deviation Principle: A belief based continuation path $\beta_i : \Delta(\Theta_{-i}) \to \mathcal{M}_i$ is optimal if it cannot be improved upon by one-shot extensions of $\beta_i(\Delta(\Theta_{-i}))$. That is, there is no $M_i \in \widetilde{\mathcal{M}}_i(\beta_i)$ and b_i such that $V_i^{M_i}(b_i) > V_i^{\beta_i(b_i)}(b_i)$.[13]

A remark is in order about the dynamics of beliefs. In general, a belief-based continuation path β_i may not be **dynamically coherent**. In other words, β_i specifies the mapping from *initial* belief (at $t = 0$) to intended path of play, and the relationship between the posterior belief in peirod $t > 0$ and the continuation path may be different form β_i. Suppose we start with initial belief b_i and an automaton $\beta_i(b_i)$, and suppose that the continuation path automaton at time t (specified by $b_i(t)$) is M_i and the posterior belief is $b_i(t)$. In general, there is no guarantee that $M_i = \beta_i(b_i(t))$ (if this is always true, we say that β_i is dynamically coherent). In other words, when the relationship between belief and continuation path at $t = 0$ is given by β_i, the relationship at time $t > 0$ may be different from β_i. Checking dynamic coherence is a demanding task. One advantage of our approach above is that the optimal belief-based continuation path β_i, identified by

---

[13] Recall that $V_i^{M_i}(b_i)$ is the expected payoff to player i under belief b_i, when player i employs on-path automaton M_i.

the simple procedure given above, automatically satisfies dynamic coherence. To state this claim formally, we need to extend our definition of belief-based optimal continuation path, because for some beliefs there could be multiple best continuation paths. So let $\overline{\beta}_i : \Delta(\Theta_{-i}) \rightrightarrows \mathcal{M}_i$ be the correspondence to represent the best continuation paths. The dynamic coherence of correspondence $\overline{\beta}_i$ is defined in an obvious way (see the proof of the following lemma). Then we have the following.

Lemma 2 *The optimal belief-based continuation path correspondence $\overline{\beta}_i$ is dynamically coherent.*

Proof. Suppose we start with initial belief b_i and an automaton $M_i \in \overline{\beta}_i(b_i)$, and suppose that, after some on-path history h_i^t the continuation path automaton is M_i' and the posterior belief is $b_i(t)$. Now suppose $\overline{\beta}_i$ is not dynamically coherent and $M_i' \notin \overline{\beta}_i(b_i(t))$. This means that M_i specifies a suboptimal actions after h_i^t, which happens with a positive probability. This contradicts the optimality of M_i. ∎

Now we compare our approach with the method proposed by Phelan and Skrzypacz (2009). Rather than focusing on path automata, they consider *extended* automata (which specify actions on *and off* the equilibrium path). They start with candidate extended preautomaton (= extended automaton without an initial state) $\overline{m}_i$, with state space Θ_i. They check the optimality of $\overline{m}_i$ by the following iterative procedure, which closely examine the dynamics of beliefs.

Phelan-Skrzypacz Procedure[14]:

1. (Step 0): Determine the set of beliefs at which one-shot deviations from extended automaton $(\overline{m}_i, \theta_i)$ are not profitable, and denote it by $Q_i^0(\theta_i)$. (**Remark:** This does not guarantee that $(\overline{m}_i, \theta_i)$ is optimal at belief $b_i \in Q_i^0(\theta_i)$, because it only guarantees that player i has no incentive to deviate *today*. It maybe the case that player i can be better off by deviating form $(\overline{m}_i, \theta_i)$ *tomorrow*. To cope with this problem, Phelan and Skrzypacz examine the dynamics of beliefs and move on to the next step.)

2. Step $k \geq 1$: Eliminate from $Q_i^{k-1}(\theta_i)$ the following beliefs b_i: there is a current action-signal pair (a_i, ω_i) for which (i) the next state is θ_i' (ii) the posterior belief is b_i' and (iii) $b_i' \notin Q_i^{k-1}(\theta_i')$. Denote the resulting set by $Q_i^k(\theta_i)$ (Remark: $Q_i^k(\theta_i)$ is the set of beliefs where deviating from extended automaton $(\overline{m}_i, \theta_i)$ at $t = 0, 2, ..., k$ is not profitable.)

3. Compute the limit $Q_i^\infty(\theta_i)$. If it is non-empty, then $(\overline{m}_i, \theta_i)$ is optimal for belief $b_i \in Q_i^\infty(\theta_i)$.

---

[14]In their notation, $Q_i^0(\theta_i) = M_i^I(\omega_i)$ and $Q_i^k(\theta_i) = (T^I)^k(M_i^I)(\omega_i)$.

In short, Phelan-Skrzypacz method recursively computes, for each step k, the set of beliefs where deviating from extended automaton $(\overline{m}_i, \theta_i)$ at $t = 0, 2, ..., k$ is not profitable. Our method based on the Belief-Based One-Shot Deviation Principle, derived from the theory of POMDP, is much simpler.

Another difference between our approach and theirs is the following. Phelan and Skrzypacz start with candidate strategies (extended preautomata), which specify behavior on and off the equilibrium path. In contrast, we start with candidate on-path behavior and use POMDP to find optimal off-path behavior. Hence their method requires an educated guess about optimal off-path behavior, while our method resorts to POMDP to find it out. Also their method describes optimal off-path behavior in terms of strategy (extended automaton), but our method describes optimal off-path behavior in terms of belief-based continuation path β_i. As Bhaskar-Obara example (2002) shows, the latter is often much simpler than the former. (The optimal belief-based continuation path in this example is simply given by a two-state path preautomata, while describing it by extended automata requires infinitely many states.)

References

Abreu, D., D. Pearce, and E. Stacchetti (1990): "Toward a Theory of Discounted Repeated Games with Imperfect Monitoring," *Econometrica,* 58, 1041-1063.

Bhaskar, V. and I. Obara (2002): "Belief-Based Equilibria in the Repeated Prisoners' Dilemma with Private Monitoring," *Journal of Economic Theory*, 102, 40-69.

Ely, J.C., J. Hörner, and W. Olszewski (2005): "Belief-free Equilibria in Repeated Games," *Econometrica*, 73, 377-415.

Ely, J.C., and J. Välimäki (2002): "A Robust Folk Theorem for the Prisoner's Dilemma," *Journal of Economic Theory*, 102, 84-105.

Fudenberg, D., D. K. Levine, and E. Maskin (1994): "The Folk Theorem with Imperfect Public Information," *Econometrica,* 62, 997-1040.

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

Hörner, J. and W. Olszewski (2006): "The Folk Theorem for Games with Private Almost-Perfect Monitoring," *Econometrica,* 74, 1499-1544.

Kaelbling, L., M. Littman, and A. Cassandra (1998): "Planning and Acting in Partially Observable Stochastic Domains," *Artificial Intelligence*, 101, 99-134.

Kandori, M. (2002): "Introduction to Repeated Games with Private Monitoring," *Journal of Economic Theory*, 102, 1-15.

Kandori, M. (2008): "Repeated Games", in *New Palgrave Dictionary of Economics, 2nd edition*, Palgrave Macmillan.

Kandori, M. and H. Matsushima (1998): "Private Observation, Communication and Collusion", *Econometrica*, 66, 627-652.

Mailath, G. and S. Morris (2002): "Repeated Games with Almost Public Monitoring", *Journal of Economic Theory*, 120, 189-228.

Mailath, G. and L. Samuelson (2006): *Repeated Games and Reputations*, Oxford University Press.

Matsushima, H. (2004): "Repeated Games with Private Monitoring: Two Players", *Econometrica*, 72, 823-852.

Obara, I. (1999): "Private Strategy and Efficiency: Repeated Partnership Game Revisited," Unpublished Manuscript, University of Pennsylvania.

Phelan, C. and A. Skrzypacz (2009) "Beliefs and Private Monitoring",
http://www.stanford.edu/~skrz/phelanskrzypacz.pdf

Piccione, M. (2002): "The Repeated Prisoner's Dilemma with Imperfect Private Monitoring," *Journal of Economic Theory*, 102, 70-83.

Sekiguchi, T. (1997): "Efficiency in the Prisoner's Dilemma with Private Monitoring", *Journal of Economic Theory*, 76, 345-361.