11-1-2009

On a Markov Game with One-Sided Incomplete Information

Johannes Hörner

Dinah Rosenberg

Eilon Solan

Nicolas Vieille

ON A MARKOV GAME WITH ONE-SIDED INCOMPLETE INFORMATION

By

Johannes Hörner, Dinah Rosenberg, Eilon Solan and Nicolas Vieille

November 2009

COWLES FOUNDATION DISCUSSION PAPER NO. 1737

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

http://cowles.econ.yale.edu/

On a Markov Game with One-Sided Information

Johannes Hörner [*], Dinah Rosenberg[†],
Eilon Solan[‡] and Nicolas Vieille[¶]

November 2, 2009

Abstract

We apply the average cost optimality equation to zero-sum Markov games, by considering a simple game with one-sided incomplete information that generalizes an example of Aumann and Maschler (1995). We determine the value and identify the optimal strategies for a range of parameters.

Keywords: repeated game with incomplete information; zero-sum games; partially observable Markov decision processes.

JEL codes: C72, C73

[*]Department of Economics, Yale University. e-mail: johannes.horner@yale.edu.

[†]Departement Economics and Decision Sciences, HEC Paris, and GREGHEC. e-mail: rosenberg@hec.fr.

[‡]The School of Mathematical Sciences, Tel Aviv University, Tel Aviv 69978, Israel. e-mail: eilons@post.tau.ac.il.

[§]Department Economics and Decision Sciences, HEC Paris, and GREGHEC. e-mail: vieille@hec.fr.

[¶]The research of Rosenberg, Solan and Vieille was supported by a grant from the Ministry of Science and Technology, Israel, and the Ministry of Research, France. The research of Solan was also supported by the Israel Science Foundation (grant No. 69/01-1).

1 Introduction

Dynamic games with incomplete information have a long history in economics. With few exceptions, these games model the uncertainty as fixed throughout the game, and private signals about the state of nature are observed only once at the beginning of the game. In many of these models, information is asymmetric, with one player's information being finer than his opponent's. Players may know their own payoffs, as in games of reputation (see Mailath and Samuelson (2006) for a survey), or they may not necessarily know those, as in the games considered by Aumann and Maschler (1968, 1995). In these games, two players are engaged in a repeated zero-sum game, whose payoff matrix depends on the state of nature that is drawn according to some commonly known distribution. The informed player observes the state; his opponent, on the other hand, has no proprietary information. Actions taken during the game are observed, but payoffs are not. Our understanding of these games spans a variety of issues: the payoffs that can be attained; the structure of the strategies achieving those; and the long-run outcome of the game.

Yet in many applications, uncertainty evolves. In reputation models, for instance, the informed player's preferences might change. Equivalently, the informed player might be replaced in a way that cannot be observed: restaurants change ownership, hedge funds change management. Such extensions are not only realistic. They are also called for by some of the more paradoxical features of traditional models. (For instance, they can explain why reputations need not disappear in the long-run.)

In economics, there is a burgeoning literature on such Markov games. See Athey and Bagwell (2008), Mailath and Samuelson (2001), Phelan (2006) and Wiseman (2008). In these papers, the *type* of one of the players is private information and affects the players' preferences over outcomes. The type changes over time according to a Markov process. On the pure game-theoretic side, Renault (2006) considers a variant of the Aumann-Maschler setup, in which the state of the game follows a Markov chain. One of the players always observes the realization of the current state, while the other player only observes the past actions chosen by the first player.

While attractive, these models remain very challenging. As optimization problems, these are partially observable competitive Markov decision processes, in which players differ in their

observation of the underlying state. Renault (2006) provides a result on the existence of the value for such games. Neyman (2008) obtains an alternative proof to Renault's result, using a reduction to the case of repeated games with incomplete information on one side à la Aumann and Maschler. However, neither paper provides a method for determining the value, or for identifying optimal strategies.

This paper illustrates the difficulties raised by such games, and suggests some methods to tackle them, within the framework of a simple, but paradigmatic example. This example has been introduced by Renault (2006), and generalizes a well-known example by Aumann and Maschler (1995). It features the basic trade-off displayed by such games, as the informed player would like to take advantage of his private information, but doing so would reveal it. Despite the simplicity of this game, we are only able to solve the game (value and optimal strategies) for a subset of parameters. We rely on the average cost optimality equation, which is well known in operations research, to verify the optimality of some strategy profiles. For the informed player, the optimal strategy that is identified is Markovian (with respect to the commonly known belief of his opponent). Remarkably, this strategy is myopically optimal in the stage game given this belief. For the uninformed player instead, the optimal strategy is represented by a finite-state automaton. The scope and generality of these properties remain to be seen. We strongly hope that our analysis piques some theorists' curiosity, and paves the way towards a more general analysis.

Our approach also suggests that numerical methods might be fruitfully applied to this type of problem, since it reduces the problem of determining the value to the study of average cost optimality equations. One of the difficulties is that the state space for one of the dynamic programming problems is continuous, namely, it is the set of possible beliefs. Polynomial methods can be used to numerically solve such problems (see, for instance, Judd (1998) for an overview), although most results are for the discounted case. One possible approach is *via* the Tauberian theorem relating the average reward criterion to the discounted reward criterion (see, for instance, Ma and Powell (2008) for a survey as well as recent results). Alternatively, there are algorithms that directly study the average optimality equation (see, for instance, Rieder (1997)).

The paper is organized as follows. Section 2 describes the game. Section 3 states the results. Section 4 gathers the proofs.

2 The game

Consider the following two-player, zero-sum, infinite-horizon game Γ_p. In any stage $n \geq 1$, the game is in one of two states, $\underline{s}$ and $\overline{s}$. The actual state s_n at stage $n \geq 1$ follows a stationary Markov chain, in which s_n is equal to s_{n-1} with probability $p \in [0,1]$. Let θ_1 denote the probability that the initial state s_1 is $\underline{s}$.

The action sets of the two players are $I = \{T, B\}$ and $J = \{L, R\}$, respectively. We denote by $g_s(i,j)$ the payoff to player 1 when the action combination (i,j) is played in state s. This payoff function is given in the two tables in Figure 1.

state $\underline{s}$

	L	R
T	1	0
B	0	0

state $\overline{s}$

	L	R
T	0	0
B	0	1

Figure 1: The payoff function.

The game is played as follows. At each stage $n \geq 1$, both players simultaneously choose actions $i_n \in \{T, B\}$ and $j_n \in \{L, R\}$. These actions are publicly observed. In addition, player 1 observes the realization of s_{n+1}. The data of the game is common knowledge among the players.

We stress that the information available to player 2 at stage n only consists in the sequence $i_1, j_1, \ldots, i_{n-1}, j_{n-1}$ of past actions, while player 1's information contains in addition the sequence $s_1, \ldots, s_{n-1}, s_n$ of states, including the current one. In particular, payoffs are not observed.[1]

The goal of player 1 is to maximize the long-run frequency of stages in which the payoff is equal to one, while the goal of player 2 is to minimize this frequency. Equivalently, the goal of player 1 is to minimize the expected time between two occurrences of payoffs of one. In other words, player 1 seeks to minimize the expected time at which his action matches both

[1]Player 1 can determine his flow payoff from his observations, while player 2 cannot. If he observed his realized payoffs, the game would become trivial.

the current state and the action of player 2. In this sense, the game Γ_p bears some similarity to so-called *rendez-vous* games, see notably Alpern and Gal (2002, 2003), Gal and Howard (2006), and Lin (2007).

Observe that, for $p = 1/2$, the states (s_n) are independent random variables. The two players then effectively face a sequence of independent, identical one-shot games with incomplete information, and the repetition of a stage-game optimal strategy is an optimal strategy in the Markov game: it is optimal to optimize payoffs myopically, because the information revealed on the current state is irrelevant for the future. At the other extreme, for $p = 1$, the current state is fixed throughout the game, and Γ_1 then coincides with the well-known Example I.2 in Aumann and Maschler (1995). In both cases, the value can be derived from the value of the one-shot game. The open and challenging question is the case $p \in (0, 1)$, which is henceforth assumed.

Since player 1 always observes the state, a naive strategy available to player 1 consists in systematically matching the current state. However, player 2 would then infer the current state from player 1's action, and would thus play next whichever action is most likely to mismatch the next period's state. Player 1 would obtain no more than $\min\{p, 1-p\}$ in the long run. As we shall see, player 1 can do significantly better.

A strategy for the uninformed player, player 2, is a function $\tau : \cup_{n\in\mathbf{N}}(I \times J)^{n-1} \to [0, 1]$, where $\tau(h)$ is the probability assigned to the action L after the sequence h of actions. A strategy for player 1 can be described by a function

$$\sigma : \cup_{n\in\mathbf{N}}(S \times I \times J)^{n-1} \to [0, 1] \times [0, 1],$$

with the following interpretation. Given a sequence h of past actions and *past* states prior to stage n, σ selects a *pair* $(x, y) = \sigma(h)$ of mixed actions. This pair specifies which mixed action is used, as a function of the current state, in stage n: the action T is played with probability x if the current state s_n turns out to be $\underline{s}$, and it is played with probability y if $s_n = \overline{s}$.

Given a strategy pair (σ, τ), let $\mathbf{P}_{\sigma,\tau}$ denote the probability measure over the set of infinite plays induced by (σ, τ), and $\mathbf{E}_{\sigma,\tau}$ the corresponding expectation operator. The distribution $\mathbf{P}_{\sigma,\tau}$ also depends on θ_1, the probability that the initial state is $\underline{s}$. Results are independent of θ_1, which is fixed throughout, and it is therefore omitted hereafter.

For a given stage $N \in \mathbf{N}$, let

$$\gamma_N(\sigma,\tau) = \frac{1}{N}\mathbf{E}_{\sigma,\tau}\left[\sum_{n=1}^{N} g_{s_n}(i_n,j_n)\right]$$

denote the average payoff in the first N stages.

A strategy σ^* of player 1 *guarantees* an amount $v \in \mathbf{R}$ if

$$\liminf_{N\to\infty} \gamma_N(\sigma^*,\tau) \geq v,$$

for every strategy τ of player 2. Player 1 *can guarantee* v if he has a strategy that guarantees v.

Similarly, a strategy τ^* of player 2 *guarantees* an amount $v \in \mathbf{R}$ if

$$\limsup_{N\to\infty} \gamma_N(\sigma,\tau^*) \leq v,$$

for every strategy σ of player 1. Player 2 *can guarantee* v if he has a strategy that guarantees v.

The real number $v \in \mathbf{R}$ is the *value* of the game if both players can guarantee v. The corresponding strategies σ^* and τ^* are then called *optimal*.

If (σ^*,τ^*) is a pair of optimal strategies, then $\gamma(\sigma^*,\tau^*) := \lim_{N\to\infty}\gamma_N(\sigma^*,\tau^*)$ does exist, and the pair (σ^*,τ^*) is a *Nash equilibrium*, in the sense that

$$\limsup_{N\to\infty} \gamma_N(\sigma,\tau^*) \leq \gamma(\sigma^*,\tau^*) \text{ and } \liminf_{N\to\infty} \gamma_N(\sigma^*,\tau) \geq \gamma(\sigma^*,\tau^*),$$

for every σ,τ. The converse also holds.

3 Results

It follows from Renault (2006), or Neyman (2008), that the game Γ_p has a value v_p, which is both the limit of the values of the N-stage games as N goes to infinity, and the limit of the values of the δ-discounted game, as the discount factor δ goes to one. As mentioned, v_p is independent of the initial distribution θ_1 (for $p \in (0,1)$). Also, because of the symmetry

of the game, it must be that $v_p = v_{1-p}$ for every p.[2] Therefore, it is sufficient to study v_p for $p \in [1/2, 1)$.

Our goal is to determine the value v_p and to identify optimal strategies for both players. We first summarize our findings regarding the value.

We are able to solve the game Γ_p for a range of values of p. We first state results concerning the value v_p.

Theorem 1 *The following holds:*

1. $v_p = \frac{p}{4p-1}$, *for* $p \in [1/2, 2/3]$, *and* $v_p \leq \frac{p}{4p-1}$ *for* $p \geq 2/3$.
2. *Let* p^* *be the unique real solution of* $9p^3 - 13p^2 + 6p - 1 = 0$ *(*$p^* \backsimeq 0.76$*). One has* $v_{p^*} = \frac{p^*}{1-3p^*+6(p^*)^2}$ *(*$v_{p^*} \backsimeq 0.35$*).*

The first statement settles an open question raised in Renault (2006) consisting in determining $v_{2/3}$. Computing the value for, say, $p = 3/4$, remains an open problem, but the statement also provides an upper bound on v_p valid for all p. A lower bound is provided below. Independently of us, and using different methods, Marino (2005) establishes that $v_p = \frac{p}{4p-1}$ for $p \in [1/2, 2/3]$. In Section 4, we describe the feature that sets apart the case $p \in [\frac{1}{2}, \frac{2}{3}]$ from the case $p > \frac{2}{3}$.[3]

We now turn to the optimal strategies that are used for $p \in [\frac{1}{2}, \frac{2}{3}]$ and $p = p^*$. Given a strategy σ of player 1, let $\theta_n(\sigma)$ denote the conditional probability that $s_n = \underline{s}$, given the actions played by both players in the first $n-1$ stages.[4] The value $\theta_n(\sigma)$ represents the belief that player 2 holds in stage n, assuming that player 1 uses strategy σ.

[2]To see this, start from a strategy σ of player 1, and define a 'mirrored' strategy σ', obtained by flipping actions and states at *even* stages. That is, given a finite history h, we first construct h' by changing at all even stages every appearance of T (resp. B, $\underline{s}$, $\overline{s}$) to B (resp. T, $\overline{s}$, $\underline{s}$), and we define $\sigma'(h)$ to be $\sigma(h')$ at odd stages, and $1 - \sigma(h')$ at even stages. Given a strategy τ for player 2, we define its mirrored version τ' in a similar way. It is immediate to check that the average payoff induced by (σ', τ') in the game Γ_{1-p} is equal to the average payoff induced by (σ, τ) in Γ. This readily implies our claim.

[3]Marino manages to compute the limit of the values of the finitely repeated game. His approach does not yield optimal strategies.

[4]Even if transition probabilities do not depend on player 2's actions, and player 2 does not observe states, the conditional probability $\theta_n(\sigma)$ may still depend on player 2's actions, since player 1's actions may depend on player 2's past actions. That is, the value of $\theta_n(\sigma)$ is independent of the strategy of player 2, but may depend on past actions of player 2, and the distribution of $\theta_n(\sigma)$ may depend on the strategy of player 2.

When checking for optimality properties of a given strategy σ for player 1, we assume the use of such a strategy, and assume that it is known to player 2, so that player 2 can recursively compute $\theta_n(\sigma)$ and act upon it. It is therefore natural to look for optimal strategies σ of player 1 in which player 1's mixed action in stage n only depends on $\theta_n(\sigma)$, and on the current state.[5]

Player 1's strategy σ^* is defined as follows. Given $\theta := \theta_n(\sigma^*)$ at stage n, the strategy σ^* prescribes the mixed action

$$\theta > \frac{1}{2} \qquad \begin{cases} \text{If } s_n = \underline{s}, & \text{Play } T \text{ with probability } \frac{1-\theta}{\theta}, \\ \text{If } s_n = \overline{s}, & \text{Play } T \text{ with probability } 0. \end{cases}$$

$$\theta \leq \frac{1}{2} \qquad \begin{cases} \text{If } s_n = \underline{s}, & \text{Play } T \text{ with probability } 1, \\ \text{If } s_n = \overline{s}, & \text{Play } T \text{ with probability } 1 - \frac{1-\theta}{\theta}. \end{cases}$$

Table 1: the strategy σ^*

We shall prove that this strategy is optimal both for $p \in [\frac{1}{2}, \frac{2}{3}]$ and for $p = p^*$. Numerical calculations seem to suggest that it is optimal for all values in the whole interval $(\frac{2}{3}, p^*)$ as well, but we have been unable to prove this. Part of the difficulty lies in the following observation. When player 1 is known to follow the strategy σ^*, the sequence of posterior beleifs held by player 2 follows a dynamical system. This dynamical system admits a Markov partition when $p \in [\frac{1}{2}, \frac{2}{3}]$, but not when $p > \frac{2}{3}$. As a consequence, it is not clear how to compute the highest amount that is guaranteed by σ^*, except for specific values of the parameter such as p^* (see Remark 6 for an elaboration on this point).

Player 2's optimal strategy differs according to the parameter range considered. We describe a first strategy τ^* of player 2, that can be implemented using a simple two-state automaton. The two states are labelled ξ_0, ξ_1. The initial state of the automaton is irrelevant. In state ξ_0 (resp. in state ξ_1), the automaton plays R (resp. L) with probability $\frac{2p}{4p-1}$. Transitions between ξ_0 and ξ_1 depend only on player 1's action and are given by:

$$(\xi_0, T) \to \xi_1, \qquad (\xi_1, T) \to \xi_1,$$
$$(\xi_0, B) \to \xi_0, \qquad (\xi_1, B) \to \xi_0.$$

That is, the automaton moves to ξ_0 (resp. to ξ_1) whenever B (resp. T) is played.

[5]There is no circularity here, since the computation of θ_n involves only the strategy of player 1 in the first $n-1$ stages.

Graphically, τ^* looks as follows.

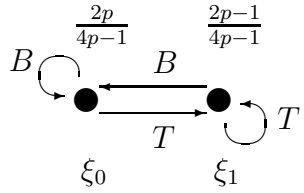

Figure 2: Player 2's strategy τ^*

In Figure 2 (and Figure 3 below), each state is labelled (below) by its name, and (above) by the probability that the action R is played in that state. Transitions are described by arrows.

For $p = p^*$, we must instead consider player 2's strategy τ^{**} that can be implemented by the following automaton with four states, labelled $1-p$, $1-\tilde{\xi}$, $\tilde{\xi}$ and p.

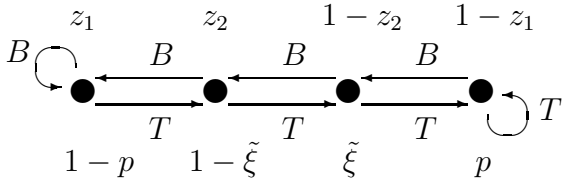

Figure 3: Player 2's strategy τ^{**}

Theorem 2 *The following holds:*

1. *The strategy σ^* is optimal for all $p \in [1/2, 2/3]$, and for $p = p^*$.*

2. *The strategy τ^* is optimal for all $p \in [1/2, 2/3]$, and the strategy τ^{**} is optimal for p^*.*

We comment on the first statement. Whenever p is not equal to $1/2$, the sequence (s_n) is autocorrelated, so that any information on s_n revealed by player 1 at stage n can be used by player 2. Casual intuition suggests that player 1 should therefore trade off the flow payoff that he obtains, with the amount of information that his action discloses about the state. Somewhat surprisingly, this trade-off is resolved in an extreme way for $p \in [1/2, 2/3]$. Indeed,

it is simple to check that the mixed action specified by σ^* at any stage n, as given in Table 1, forms an optimal mixed action of player 1 in the *one-shot* game with incomplete information in which the state is drawn with probabilities $\theta_n(\sigma^*)$ and $1-\theta_n(\sigma^*)$, and only player 1 is informed of the state. In this sense, the strategy σ^* is *myopically* optimal.[6] Thus, statement 1 implies that when the underlying Markov chain is sufficiently mixing, playing in a myopically optimal way is optimal in the long run for such values of p. It can be shown that a similar result holds in any Markov game.[7]

According to the second statement, an optimal strategy of player 2 can be found within the strategies that can be implemented with finite automata. Whether this holds for all parameters remains an open question.

Let us now briefly discuss the case of discounted versions of the game. Our proof follows a *guess-and-check* approach: we guess optimal strategies and check optimality using a verification theorem. We could use a similar approach and analyze discounted games, using a dynamic programming principle. Note that both the value and the optimal strategies would then depend on the distribution of the initial state.

It is plausible that σ^* is an optimal strategy in discounted games as well. Indeed, if a myopic strategy is optimal when payoffs are not discounted, it might *a fortiori* be optimal when some positive weight is assigned to current payoffs. On the other hand however, none of the strategies τ^* and τ^{**} can possibly be an optimal strategy in discounted versions of the game, except possibly for highly specific initial distributions of the state. We have no guess to offer on the optimal behavior of player 2 in discounted games.

In spite of this negative comment, Theorem 2 yields some results in discounted games. It can be checked that the strategy σ^* is approximately optimal in all long, finitely repeated games:[8] given $\varepsilon > 0$, there is an horizon N such that $\gamma_n(\sigma^*, \tau) \geq v_p - \varepsilon$ for every strategy τ, whenever the length n of the game exceeds N. Together with the property that the values of finitely repeated games and of discounted games here converge to the same value, this implies that σ^* is an ε-optimal strategy in all discounted games, provided the discount factor is high

[6] Indeed, σ^* plays at every stage as if the current stage were the last one.

[7] Note that the mixed action in Table 1 is not the unique optimal strategy in the one-shot game. However, among optimal mixed actions, it is the only one for which player 2 is indifferent between playing L and R, and it is the one that reveals the least amount of information about the state.

[8] Similar claims hold for τ^* and τ^{**}.

enough.

As mentioned, we have not established that σ^* is optimal for all values in the range $(2/3, p^*)$, although it appears to be so numerically. As it turns out (see Section 4), the payoff that player 1 obtains by following this strategy, denoted by γ_p, is independent of the strategy used by player 2. Therefore, it provides a lower bound to the value: $v_p \geq \gamma_p$. We show in the next section that

$$\frac{1}{\gamma_p} = u_0 + u_0 u_1 + u_0 u_1 u_2 + \cdots,$$

where the sequence (u_n) is defined by $u_0 = 1$ and, for $\psi(u) := 3p - 1 - \dfrac{2p-1}{u}$,

$$u_{n+1} = \max\{\psi(u_n), 1 - \psi(u_n)\}.$$

There seems to be no explicit solution to this sequence, except for a countable set of points, as explained and described in Subsection 4.4.

Figure 1 features both the upper bound $p/(4p-1)$, the lower bound γ_p and, in between, the value v_p. The latter two functions are obtained numerically.

4 Proofs

Proofs are organized as follows. In Subsection 4.1, we state the main theorem from Markov decision processes that is used throughout, and explain how to apply it to a Markov game. In Subsections 4.2 and 4.3, we prove all claims relative to players 2 and 1 respectively. Subsection 4.4 provides a lower bound on v_p.

4.1 Average Cost Optimality Equations and Applications

We here state a version of the well-known Average Cost Optimality Equation (ACOE) for general Markov Decision Processes (MDP). We next show how to use it to prove the optimality of specific simple strategies in Markov games with incomplete information, such as Γ_p.

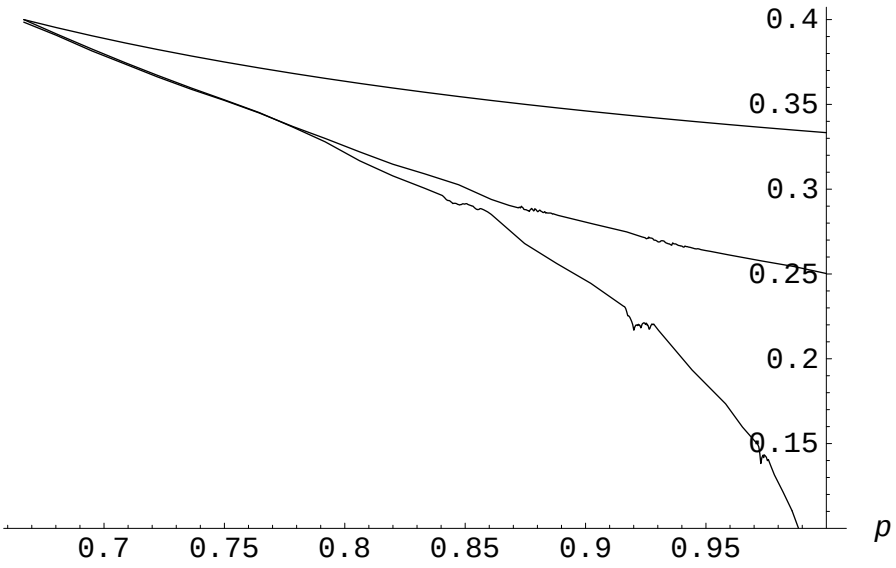

Figure 1: The upper bound $p/(4p-1)$, the value vp and the lower bound γ_p.

Proposition 3 *Let $(S, \mathcal{A}, r, q)$ be a MDP with a compact metric space S, a compact action set $\mathcal{A}$, a continuous payoff function $r : S \times \mathcal{A} \to \mathbf{R}$, and a continuous transition rule $q\colon S \times \mathcal{A} \to \Delta(S)$ such that $q(\cdot \mid s, a)$ has finite support for every $(s, a) \in S \times \mathcal{A}$.*

If there is $v \in \mathbf{R}$ and a bounded function $V : S \to \mathbf{R}$ such that

$$v + V(s) = \max_{a \in \mathcal{A}} \left(r(s,a) + \sum_{s' \in S} q(s' \mid s, a) V(s') \right) \text{ for each } s \in S, \tag{1}$$

then v is the value of the MDP, for each initial state $s \in S$. Furthermore, a stationary strategy $\alpha = (\alpha(s))$ is optimal whenever $\alpha(s)$ attains the maximum in the right-hand side of (1), for every $s \in S$.

Moreover, if, there is $s^* \in S$ such that the sequence $n\,(v_n(s) - v_n(s^*))$ has a point-wise limit V for every $s \in S$, and if the value v of the MDP is independent of the initial state, then

$$v + V(s) = \max_{a \in \mathcal{A}} \left(r(s,a) + \sum_{s' \in S} q(s' \mid s, a) V(s') \right) \text{ for each } s \in S. \tag{2}$$

For a proof, see, e.g., Feinberg and Shwartz (2002). In Proposition 3, $\Delta(S)$ stands for the set of probability distributions over S. The value v is the supremum over all policies ϕ

of the payoff function $\gamma(\phi) := \liminf_{N\to\infty} \mathbf{E}_\phi \left[\frac{1}{N}\sum_{n=1}^{N} r(s_n, a_n)\right]$. Observe that the function V is determined only up to an additive constant.

Consider a Markov game with incomplete information, such as the game Γ_p. Suppose first that one is looking for the highest long-run payoff v that player 1 can guarantee when using σ^*. The strategy σ^* is given, so that player 2 effectively faces a MDP over the state space $[0,1]$, where an element $\theta \in [0,1]$ is interpreted as the conditional probability $\theta_n(\sigma^*)$ that player 2 attaches to the state being $\underline{s}$. The value of this MDP coincides with the highest long-run payoff v that σ^* guarantees.

It turns out that in the game Γ_p, the strategy σ^* that was defined in Table 1 satisfies the following: player 2 is always indifferent between playing L and R. Hence, the value of this MDP can also be computed by restricting player 2 to one action, say L. In particular, letting r and q denote the reward and the transition functions of this auxiliary MDP, v is characterized by the existence of a function V over $[0,1]$ such that

$$v + V(\theta) = r(\theta, L) + \sum_{\theta' \in S} q(\theta' \mid \theta, L) V(\theta') \text{ for each } \theta.$$

On the other hand, let now τ be a strategy that can be implemented by a finite automaton (with deterministic transitions). Suppose that one is looking for the highest long-run payoff v that player 2 can guarantee when using τ. Since player 1 observes the actions chosen by both players, he can compute at every stage the current state of the automaton. Since he also knows the current state of the game, player 1 essentially is facing a MDP whose state space is the product of the automaton's state space and of $\{\underline{s}, \overline{s}\}$. The ACOE can be used to find the value of the MDP, hence to find v, and a best reply to τ.

4.2 Player 2

Here, we prove that τ^* guarantees $\frac{p}{4p-1}$ for all $p \geq \frac{1}{2}$, which establishes the first part of Theorem 1.

Lemma 4 *The strategy τ^* of player 2 guarantees $p/(4p-1)$ for all $p \geq 1/2$.*

Proof. As player 1 knows the state of player 2's automaton and the current state, he essentially faces a MDP with four states: $\mathcal{S} = \{(\underline{s}, \underline{\xi}), (\bar{s}, \underline{\xi}), (\underline{s}, \bar{\xi}), (\bar{s}, \bar{\xi})\}$. In each one of these four states he has two available actions, T and B, and the payoff is

$$\begin{array}{ll} r((\underline{s}, \underline{\xi}); T) = \frac{2p}{4p-1} & r((\underline{s}, \underline{\xi}); B) = 0, \\ r((\bar{s}, \underline{\xi}); T) = 0 & r((\bar{s}, \underline{\xi}); B) = \frac{2p-1}{4p-1}, \\ r((\underline{s}, \bar{\xi}); T) = \frac{2p-1}{4p-1} & r((\underline{s}, \bar{\xi}); B) = 0, \\ r((\bar{s}, \bar{\xi}); T) = 0 & r((\bar{s}, \bar{\xi}); B) = \frac{2p}{4p-1}. \end{array}$$

Consider now the transition out of state $(\underline{s}, \underline{\xi})$, when player 1 plays T. Following T, the automaton state moves to $\bar{\xi}$. On the other hand, the state of the game moves to $\bar{s}$ with probability p. As a result, the next state of the MDP is $(\underline{s}, \bar{\xi})$ with probability $1-p$, and $(\bar{s}, \bar{\xi})$ otherwise. All other transitions are obtained in the same fashion.

By the ACOE, $v = \frac{p}{4p-1}$ is the value if and only if there is a function $V : \mathcal{S} \to \mathbf{R}$ that satisfies

$$\begin{aligned} v + V(\underline{s}, \underline{\xi}) = \max \Bigg\{ & \frac{2p}{4p-1} + pV(\underline{s}, \bar{\xi}) + (1-p)V(\bar{s}, \bar{\xi}), \\ & pV(\underline{s}, \underline{\xi}) + (1-p)V(\bar{s}, \underline{\xi}) \Big\}, \end{aligned} \tag{3}$$

$$\begin{aligned} v + V(\bar{s}, \underline{\xi}) = \max \Big\{ & pV(\bar{s}, \bar{\xi}) + (1-p)V(\underline{s}, \bar{\xi}), \\ & \frac{2p-1}{4p-1} + pV(\bar{s}, \underline{\xi}) + (1-p)V(\underline{s}, \underline{\xi}) \Bigg\}, \end{aligned} \tag{4}$$

$$\begin{aligned} v + V(\bar{s}, \bar{\xi}) = \max \Bigg\{ & \frac{2p}{4p-1} + pV(\bar{s}, \underline{\xi}) + (1-p)V(\underline{s}, \underline{\xi}), \\ & pV(\bar{s}, \bar{\xi}) + (1-p)V(\underline{s}, \bar{\xi}) \Big\}, \end{aligned} \tag{5}$$

$$\begin{aligned} v + V(\underline{s}, \bar{\xi}) = \max \Big\{ & pV(\underline{s}, \underline{\xi}) + (1-p)V(\bar{s}, \underline{\xi}), \\ & \frac{2p-1}{4p-1} + pV(\underline{s}, \bar{\xi}) + (1-p)V(\bar{s}, \bar{\xi}) \Bigg\}. \end{aligned} \tag{6}$$

These equations are symmetric. As is easy to check, they imply that $V(\underline{s},\bar{\xi}) = V(\bar{s},\underline{\xi})$ and $V(\bar{s},\bar{\xi}) = V(\underline{s},\underline{\xi})$, so Eqs. (5) and (6) can be omitted. One solution to these equations is

$$\begin{aligned} v &= \frac{p}{4p-1}, \\ V(\underline{s},\underline{\xi}) &= V(\bar{s},\bar{\xi}) = 0, \\ V(\bar{s},\underline{\xi}) &= V(\underline{s},\bar{\xi}) = -\frac{1}{4p-1}. \end{aligned}$$

■

4.3 Player 1

Here, we prove the results concerning player 1. First, we will prove that σ^* guarantees $p/(4p-1)$ whenever $p \in [1/2, 2/3]$. Together with the results of the previous subsection, this yields $v_p = p/(4p-1)$, and implies that σ^* and τ^* are indeed optimal for that range of values of p.

Next, we will consider the case where $p = p^*$, and prove that σ^* and τ^{**} are optimal strategies.

4.3.1 The case $p \in [1/2, 2/3]$

We here prove that σ^* is optimal in Γ_p, for all $p \in [1/2, 2/3]$. Under σ^*, and when the posterior probability of $\underline{s}$ is $\theta < \frac{1}{2}$, player 1 plays as follows:

State $\underline{s}$

	L	R
1	1	0
0	0	0

$\mathbf{P}(s_n = \underline{s}) = \theta_n(\sigma^*) = \theta$

State $\overline{s}$

	L	R
$1 - \frac{\theta}{1-\theta}$	0	0
$\frac{\theta}{1-\theta}$	0	1

$\mathbf{P}(s_n = \bar{s}) = 1 - \theta_n(\sigma^*) = 1 - \theta$

Observe that the total probability that the action B is played is θ, so that the total probability that the action T is played is $1 - \theta$.

Lemma 5 *The strategy σ^* guarantees $p/(4p-1)$, for each $p \in [1/2, 2/3]$.*

Proof. Consider the MDP over the state space $[0,1]$ of posterior beliefs, induced by σ^*. Recall that the action of player 2 does not affect the evolution of the posterior belief and note that, when facing σ^*, both actions L and R yield the same current payoff, θ. Therefore, the value of this MDP can be found by restricting player 2 to play, say, L in each and every stage.

Transitions in this MDP are as follows. For concreteness, let $\theta \leq 1/2$:

- If player 1 played B at some stage, the state at that stage was $\bar{s}$. Therefore, player 2's posterior belief assigns probability $1-p$ to the next state being $\underline{s}$;
- If player 1 played T at some stage, player 2 updates his belief, and assigns a probability $\theta/(1-\theta)$ to the current state being $\underline{s}$. Therefore, player 2's posterior belief assigns probability
$$\frac{\theta}{1-\theta}p + (1-p)\left(1 - \frac{\theta}{1-\theta}\right) = 1 - p + (2p-1)\frac{\theta}{1-\theta}$$
to the next state being $\underline{s}$.

In order to prove that the strategy σ^* guarantees $p/(4p-1)$, we therefore need to find a function V, symmetric around $1/2$, such that the equality

$$\frac{p}{4p-1} + V(\theta) = \theta + (1-\theta)V\left(1 - p + (2p-1)\frac{\theta}{1-\theta}\right) + \theta V(1-p) \tag{7}$$

holds for each $\theta \leq 1/2$.

Observe that one has $1 - p + (2p-1)\frac{\theta}{1-\theta} \geq 1/2$, whenever $p \in [1/2, 2/3]$: under σ^*, the posterior probability is either p, $1-p$, or jumps from one half of the posterior space $[1-p, p]$ to the other.

Therefore, we need only find a function $V : [1-p, 1/2] \to \mathbf{R}$, such that

$$\frac{p}{4p-1} + V(\theta) = \theta + (1-\theta)V\left(p - (2p-1)\frac{\theta}{1-\theta}\right) + \theta V(1-p).$$

One can verify that the function $V(\theta) = \dfrac{\theta}{4p-1}$ is a solution. The result follows. ■

Remark 6

There is little hope to extend this proof for values of p beyond $2/3$. Indeed, the above proof rests on the fact that (7) has a simple solution, V. This in turn is related to the fact that, given θ, the belief $1-p+(2p-1)\frac{\theta}{1-\theta}$ always lies on the opposite side of $1/2$. That is, leaving aside possible transitions to p or $1-p$, the dynamics of beliefs admits a Markov partition which consists of two intervals: $[1-p, 1/2]$ is mapped into $[1/2, p]$, and vice-versa. As a result, V has a piecewise linear structure, which consists of two parts. It can be shown that, for $p > 2/3$, the corresponding dynamics admits no finite Markov partition.

4.3.2 The case $p = p^*$

We here analyze the case where $p = p^*$, the unique real solution to the equation $9p^3 - 13p^2 + 6p - 1 = 0$.

We establish that $v_p = \frac{p}{1-3p+6p^2}$ by showing that it is an equilibrium payoff. Since the game is zero-sum, the equilibrium consists of optimal strategies. We are going to show that the following strategy profile is an equilibrium: player 1 uses strategy σ^*, as defined in Table 1, and player 2 uses strategy τ^{**}, as defined in Figure 3.

When facing σ^*, player 2 is always indifferent between playing L or R. Since σ^* is independent of player 2's actions, the payoff is independent of player 2's strategy. In particular, player 2's automaton is a best-reply to σ^*.

We now prove that σ^* is a best-reply to τ^{**}, for appropriate values of z_1 and z_2. As before, player 1 is facing an MDP with 8 states. Each state is composed of the current state (2 alternatives) and the state of player 2's automaton (4 alternatives).

View the labels $1-p, 1-\tilde{\xi}, \tilde{\xi}, p$ of player 2's automaton as *fictitious* beliefs, set $\tilde{\xi} = \frac{p}{3p-1} \backsimeq 0.5944$, and let θ_1, the distribution of the initial state, coincide with the initial state of the automaton.[9]

We first claim that, for $p = p^*$ and for our choice of $\tilde{\xi}$, the state ξ_n of player 2's automaton at stage n coincides with $\theta_n(\sigma^*)$.

Indeed, suppose that player 2 knows that he is facing σ^*, and computes beliefs accordingly. Denote by $\phi(\theta \mid a)$ the belief of player 2 in stage $n+1$, if his belief was θ in stage n, and after

[9]Recall that the distribution of the initial state is irrelevant for the determination of the value.

player 1 played $a \in \{T, B\}$ in stage n. By Bayes' rule, and for $\theta > 1/2$, one has

$$\phi(\theta \mid B) = \frac{2\theta - 1}{\theta} p + (1 - p) \left(1 - \frac{2\theta - 1}{\theta}\right) = p - (2p - 1) \times \frac{1 - \theta}{\theta},$$
$$\phi(\theta \mid T) = p.$$

Using elementary manipulations, one can verify that

$$\phi(p \mid B) = \tilde{\xi} \text{ and } \phi(\tilde{\xi} \mid B) = 1 - \tilde{\xi} \tag{8}$$

whenever player 1 plays B, the posterior belief of player 2 evolves as follows: (i) from p it moves to $\tilde{\xi}$, (ii) from $\tilde{\xi}$ it moves to $1 - \tilde{\xi}$ whereas (iii) from either $1 - \tilde{\xi}$ or $1 - p$ it moves to $1 - p$.

By symmetric arguments, this implies that the transitions of the automaton of player 2 mimic the evolution of his posterior belief, provided player 1 follows σ^*. This implies in turn that the strategy σ^* induces a *stationary Markov* strategy in the auxiliary 8-state MDP.

Recall now that the strategy σ^* assigns positive probability to both actions T and B whenever the current state is $\underline{s}$ and $\theta_n(\sigma^*) > 1/2$, and whenever the current state is $\bar{s}$ and $\theta_n(\sigma^*) < 1/2$, and assigns otherwise probability 1 to either T or B.

That is, the stationary strategy of player 1 associated with σ^* (i) assigns positive probability to both actions in the four states $(\underline{s}, \tilde{\xi})$, $(\underline{s}, p)$, $(\bar{s}, 1 - \tilde{\xi})$ and $(\bar{s}, 1 - p)$, (ii) plays T in states $(\underline{s}, 1 - \tilde{\xi})$ and $(\underline{s}, 1 - p)$ and (iii) plays B in states $(\bar{s}, p)$ and $(\bar{s}, \tilde{\xi})$.

By symmetry properties, and using the ACOE, this stationary strategy is an optimal strategy in the MDP with value v as soon as there is a function V such that the following

system of equations and inequations is satisfied:

$$v + V(\underline{s}, 1-p) = z_1 + pV(\underline{s}, 1-\tilde{\xi}) + (1-p)V(\bar{s}, 1-\tilde{\xi}), \tag{9}$$

$$v + V(\underline{s}, 1-p) \geq pV(\underline{s}, 1-p) + (1-p)V(\bar{s}, 1-p), \tag{10}$$

$$v + V(\underline{s}, 1-\tilde{\xi}) = z_2 + pV(\underline{s}, \tilde{\xi}) + (1-p)V(\bar{s}, \tilde{\xi}), \tag{11}$$

$$v + V(\underline{s}, 1-\tilde{\xi}) \geq pV(\underline{s}, 1-p) + (1-p)V(\bar{s}, 1-p), \tag{12}$$

$$v + V(\underline{s}, \tilde{\xi}) = 1 - z_2 + pV(\underline{s}, p) + (1-p)V(\bar{s}, p), \tag{13}$$

$$v + V(\underline{s}, \tilde{\xi}) = pV(\underline{s}, 1-\tilde{\xi}) + (1-p)V(\bar{s}, 1-\tilde{\xi}), \tag{14}$$

$$v + V(\underline{s}, p) = 1 - z_1 + pV(\underline{s}, p) + (1-p)V(\bar{s}, p), \tag{15}$$

$$v + V(\underline{s}, p) = pV(\underline{s}, \tilde{\xi}) + (1-p)V(\bar{s}, \tilde{\xi}). \tag{16}$$

Eqs. (9) and (10) express the fact that it must be optimal to play T when in state $(\underline{s}, 1-p)$. The right-hand side of (10) is obtained by noting that, when playing T in state $(\underline{s}, 1-p)$, the current reward to player 1 is z_1, and the MDP moves to state $(\underline{s}, 1-\tilde{\xi})$ with probability p, and to state $(\bar{s}, 1-\tilde{\xi})$ otherwise. Eq. (11) is obtained by noting that, when playing B in state $(\underline{s}, 1-p)$, the current reward to player 1 is 0, and the MDP moves to state $(\underline{s}, 1-p)$ with probability p, and to state $(\bar{s}, 1-p)$ otherwise. The other conditions are obtained in a similar fashion.

Since V is determined up to an additive constant, we can set $V(\underline{s}, p) = 0$, and then this system contains 6 equations in 6 variables. The unique solution is

$$\begin{aligned}
v &= \frac{p}{1-3p+6p^2} \backsimeq 0.348291466, \\
z_1 &= \frac{4p-1}{1-3p+6p^2} \backsimeq 0.934232129, \\
z_2 &= \frac{2p}{1-3p+6p^2} \backsimeq 0.696582932, \\
V(\underline{s}, 1-p) &= \frac{6p-2}{1-3p+6p^2} \backsimeq 1.171881327, \\
V(\underline{s}, 1-\theta) &= \frac{2p}{1-3p+6p^2} \backsimeq 0.696582932, \\
V(\underline{s}, \theta) &= \frac{2p-1}{1-3p+6p^2} \backsimeq 0.237649197, \\
V(\underline{s}, p) &= 0.
\end{aligned}$$

One can verify that the inequalities (10) and (12) are satisfied in this case.

4.4 A lower bound on v_p

Recall that γ_p is the payoff that player 1 obtains if he uses strategy σ^*, which is independent of the strategy used by player 2. Plainly $v_p \geq \gamma_p$. In the present subsection we compute γ_p, thereby providing a lower bound to v_p.

By the ACOE, the number γ_p is the unique real number such that there is a function V (symmetric around 1/2) that satisfies

$$\gamma_p + V(\theta) = \theta + (1-\theta)V\left(1-p+(2p-1)\frac{\theta}{1-\theta}\right) + \theta V(1-p) \qquad \theta \leq 1/2,$$
$$\gamma_p + V(\theta) = 1-\theta + \theta V\left(p-(2p-1)\frac{1-\theta}{\theta}\right) + (1-\theta)V(p) \qquad \theta \geq 1/2.$$

As V can be determined up to an additive constant, set $V(1-p)=0$. Set $\theta_0 = 1-p$, and for each $k \in \mathbf{N}$ set

$$\theta_{k+1} = \min\left\{1-p+(2p-1)\frac{\theta_k}{1-\theta_k}, p-(2p-1)\frac{\theta_k}{1-\theta_k}\right\}.$$

Then $\theta_k \in [1-p, 1/2]$, and

$$\gamma_p + V(\theta_k) = \theta_k + (1-\theta_k)V(\theta_{k+1}) + \theta_k V(1-p).$$

Using $V(1-p)=0$, and given the boundedness of V, it follows that

$$\gamma_p = \frac{\theta_0 + (1-\theta_0)\theta_1 + (1-\theta_0)(1-\theta_1)\theta_2 + \cdots}{1+(1-\theta_0)+(1-\theta_0)(1-\theta_1)+\cdots}.$$

The sequence (u_n) was defined by $u_0 = 1$ and $u_{n+1} = 1-\theta_n$, so that (u_n) satisfies the recursive equation

$$u_{n+1} = \max\left\{3p-1-\frac{2p-1}{u_n}, 2-3p+\frac{2p-1}{u_n}\right\}.$$

The payoff γ_p is then given by

$$\begin{aligned}\frac{1}{\gamma_p} &= \frac{u_0 + u_0u_1 + u_0u_1u_2 + \cdots}{1-u_1+u_1(1-u_2)+u_1u_2(1-u_3)+\cdots} \\ &= u_0 + u_0u_1 + u_0u_1u_2 + \cdots\end{aligned}$$

Observe further that, as mentioned, the recurrence equation on (u_n) writes

$$u_{n+1} = \max\{\psi(u_n), 1 - \psi(u_n)\},$$

where $\psi(u) := 3p - 1 - \dfrac{2p-1}{u}$ is increasing. Let u^* be a solution to $u = 1 - \psi(u)$. It is immediate to check that $u^* \geq 1/2$, hence $\psi(u^*) \leq u^*$, for otherwise, the inequality $2u^* < \psi(u^*) + 1 - \psi(u^*) = 1$ would hold. In particular, if $u_N = u^*$ for some N, then the sequence (u_n) is stationary from that stage on.

Next, consider the sequence (w_n) defined by $w_0 = 1 = u_0$ and

$$w_{n+1} = 3p - 1 - \frac{2p-1}{w_n}. \tag{17}$$

We claim that if $w_N = u^*$ (and $w_n \neq u^*$ for $n < N$), then $u_n = w_n$ for each $n < N$ and $u_n = u^*$ for each $n \geq N$. To prove this claim, it is enough to check that $1 - \psi(w_n) \leq \psi(w_n)$ for $n = 1, 2, \cdots, N-1$. To see why this holds, observe first that the sequence (w_n) is decreasing. We argue by contradiction, and assume that $w_{k+1} = \psi(w_k) < 1 - \psi(w_k)$ for some $k < N$. Since $w_k > w_{k+1} \geq w_N$, this yields $w_N < 1 - \psi(w_k)$. Since ψ is increasing, one obtains $w_N < 1 - \psi(w_N)$ – a contradiction.

As a result, the computation of γ_p is easy for those values of p with the property that $w_n = u^*$ for some n.

The solution to (17) is

$$w_n = \sqrt{2p-1}\frac{\cos\left((n+1)\,\sigma - \lambda\right)}{\cos\left(n\sigma - \lambda\right)}, \tag{18}$$

where

$$\tan\sigma = \frac{\sqrt{(1-p)\,(9p-5)}}{(3p-1)}, \tan\lambda = \frac{3(1-p)}{\sqrt{(1-p)\,(9p-5)}}.$$

Using standard manipulations, the following appears. Given N, there exists a unique p such that N is the smallest integer for which $u_{N-1} = u_N$. This p solves the polynomial equation

$$\tan((N+1)\sigma) = \sqrt{\frac{9p-5}{1-p}} \times \frac{2p-1}{1-3p+\sqrt{p(9p-4)}}.$$

For instance, (i) $p \backsimeq .7589$ for $N = 2$, which is the special case already studied, (ii) $p \backsimeq .8583$ for $N = 3$, (iii) $p \backsimeq .9073$ for $N = 4$, (iv) $p \backsimeq .9348$ for $N = 5$, etc.

From the expression of w_n, one deduces

$$w_0 \cdots w_n = (2p-1)^{(n+1)/2} \frac{\cos\left((n+1)\,\sigma - \lambda\right)}{\cos\lambda}.$$

It follows that

$$\frac{1}{\gamma_p} = 2\left(1 - \frac{(2p-1)^{(N+1)/2}\sqrt{2}(3p-1-\sqrt{p(9p-4)})}{\sqrt{p(1-p)}(3\sqrt{p}-\sqrt{9p-4})(5p-2-\sqrt{p(9p-4)})}\right).$$

For instance, $\gamma_p \backsimeq .2880$ for $p \backsimeq .8583$ (more precisely, γ_p is the unique real root of $-162 + 1737x - 7279x^2 + 15002x^3 - 15276x^4 + 6169x^5 = 0$ for p the unique real root of $-4 + 37x - 136x^2 + 248x^3 - 225x^4 + 81x^5 = 0$), $\gamma_p \backsimeq .2460$ for $p \backsimeq .9073$, etc. It is readily verified that the equation $\gamma_p = \frac{p}{1-3p+6p^2}$, valid for $p \backsimeq .7589$, is not valid on any open interval around this p.

References

[1] Alpern, S. and S. Gal (2002), Searching for an Agent who may or may not Want to be Found, *Operations Research*, **50**, 311–323.

[2] Alpern, S. and S. Gal (2003), *The Theory of Search Games and Rendezvous*, International Series in Operations Research and Management Science, Volume 55, Kluwer Academic Publishers, Boston.

[3] Athey, S. and K. Bagwell (2008), Collusion with Persistent Cost Shocks, *Econometrica*, **76**(3), 493–540.

[4] Aumann, R.J. and M.B. Maschler (1995), *Repeated Games with Incomplete Information*, The MIT Press.

[5] Feinberg, E.A. and Shwartz A. (2002), *Handbook of Markov Decision Processes*, Kluwer.

[6] Gal, S. and J.V. Howard (2006), Rendezvous-Evasion Search in Two Boxes, *Operations Research*, **53**, 689–697.

[7] Judd, K. L. (1998), *Numerical Methods in Economics*, MIT Press, Cambridge.

[8] Lim, W.S. (1997), Rendezvous-Evasion Games on Discrete Locations, with Joint Randomization, *Advances in Applied Probability*, **29**, 1004–1017.

[9] Ma, J. and W. B. Powell (2008), "Convergence Proofs of Least Squares Policy Iteration Algorithm for High-Dimensional Infinite Horizon Markov Decision Process Problems," working paper, Princeton University.

[10] Mailath, G. and L. Samuelson (2001), Who Wants a Good Reputation?, *Review of Economic Studies,* **68**, 415–441.

[11] Mailath, G. and L. Samuelson (2006), *Repeated Games and Reputations: Long-Run Relationships*, Oxford University Press.

[12] Marino, A. (2005), The Value and Optimal Strategies of a Particular Markov Chain Game. Preprint.

[13] Neyman, A. (2008), Existence of Optimal Strategies in Markov Games with Incomplete Information, *International Journal of Game Theory*, **37**, 518–596.

[14] Phelan, C. (2006), Public Trust and Government Betrayal, *Journal of Economic Theory*, **127**(1), 27–43.

[15] Renault, J. (2006), The Value of Markov Chain Games with Lack of Information on One Side, *Mathematics of Operations Research*, **31**, 490–512.

[16] Rieder , U. (1997), "Average Optimality in Markov Games with General State Space," *Proceedings of the 3rd International Conference on Approximation and Optimization in the Caribbean.*

[17] Wiseman, T. (2008), Reputation and Impermanent Types, *Games and Economic Behavior*, **62**, 190–210.