Ann Oper Res
DOI 10.1007/s10479-015-1905-6

Value of information for a leader–follower partially observed Markov game

Yanling Chang[1] · **Alan L. Erera**[1] · **Chelsea C. White III**[1]

Abstract We consider a leader–follower partially observed Markov game (POMG) and analyze how the value of the leader's criterion changes due to changes in the leader's quality of observation of the follower. We give conditions that insure improved observation quality will improve the leader's value function, assuming that changes in the observation quality do not cause the follower to change its policy. We show that discontinuities in the value of the leader's criterion, as a function of observation quality, can occur when the change of observation quality is significant enough for the follower to change its policy. We present conditions that determine when a discontinuity may occur and conditions that guarantee a discontinuity will not degrade the leader's performance. We show that when the leader and the follower are collaborative and the follower completely observes the leader's initial state, discontinuities in the leader's value function will not occur. However, examples show that improving observation quality does not necessarily improve the leader's criterion value, whether or not the POMG is a collaborative game.

Keywords Dynamic programming · Artificial intelligence · Sequential decision making

1 Introduction

The research in this paper considers the following multi-period stochastic game. There are two intelligent and adaptive decision-making agents, a leader and a follower. These agents interact as follows:

✉ Yanling Chang
changyanling@gatech.edu

Alan L. Erera
alerera@isye.gatech.edu

Chelsea C. White III
chip.white@isye.gatech.edu

[1] H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA

Published online: 27 May 2015

- Before the game begins, the follower chooses its policy with complete knowledge of what policy the leader has chosen.
- Once the game begins, the policies chosen by the leader and follower determine what actions are simultaneously selected at each decision epoch of an at most countable number of decision epochs.
- The decisions of both policies affect the dynamics of the system that both agents want to control.
- Each agent's policy makes decisions to achieve its agent's objective, based on data that include possibly inaccurate, incomplete, and/or costly observations of the other agent.

The question addressed in this paper is: how does the accuracy of the leader's observation of the follower affect the performance of the leader? A related question: what is the added value to the leader if more accurate (and presumably, requiring greater resources) observations of the follower could be made available? Another related question: is it guaranteed that more accurate information about the follower will improve leader performance? We use a leader–follower partially observed Markov game (POMG) to model and address these "value of information" questions by building on previous research presented in (Chang et al. 2014), specialized to the case where the leader has a single objective [(the leader was assumed to have multi-objectives in (Chang et al. 2014)].

It seems intuitive that more accurate observations of the underlying state of a system subject to control will lead to better system performance. Better medical diagnostic quality should lead to healthier patient outcomes; higher quality machine fault identification should lead to more effective machine maintenance and better system performance; more accurate observations of an adversary or a competitor should provide advantage. However, several counterexamples to this claim for the partially observed Markov decision process (POMDP), a single-agent decision making model, have been found in Ortiz et al. (2013).

The assumption that the follower has complete knowledge of the leader's policy is realistic in many applications and can serve as a worst-case scenario ("erring" in the right direction) otherwise. With respect to realism, if the leader is defending a key infrastructure site and the follower's objective involves breaching site security, the follower may have at least partial knowledge of the leader's policy by observing the leader's allocation of defensive resources. Also, the follower may be able to infer, at least in part, the leader's policy from sources such as publicly available information if, for example, the leader is associated with a government agency with open records. We remark that the underlining state of each agent is only partially observed by the other agent and hence each agent's policy selects actions based on data different from the other agent's data. Thus, in general the follower will not know exactly what action the leader will take, which may mollify concern in situations where the leader–follower assumption is believed to unrealistically bias results to the advantage of the follower.

The intent of the research is to better inform the decision to seek or not seek improved state observation quality and what resulting changes in performance to expect if state observation quality is improved. The initial motivation for this research was how to support a manager of a food supply chain (the leader) in selecting a sequence of actions that best balances maximizing the performance of, while minimizing the risk to, the supply chain over time, given that there is an attacker (the follower) who seeks to contaminate the food supply chain with a chemical or biological toxin. More specifically in the context of this application, if the manager (or a government agency) is able to improve the quality of observing the attacker (presumably at a cost), is it useful to do so? Details of this application can be found in Chang et al. (2014). A detailed discussion of the reasons behind the interest in measures of risk and risk mitigation when an adversary is intelligent and adaptive can be found in Ezell et al. (2010).

This paper is organized as follows. Section 2 presents a review of the literature. We consider the single agent case in Sect. 3. This case and its associated value determination are important for results to follow and are the basis of the solution procedure for the POMG found in (Chang et al. 2014) because:

- Given a leader policy, the (two-agent) POMG can be transformed into a (single agent, potentially tractable) POMDP that determines the follower's response policy.
- Computing the value of either the leader's objective function or the follower's objective function requires both a leader's policy and a follower's policy.

The POMDP problem statement and preliminary results from the literature are presented in Sect. 3.1 through 3.3. We summarize two definitions of observation quality and related results from the literature. Theorem 1 and Corollary 1 present conditions that guarantee that improved observation quality insures improved system performance for the first definition. Results with respect to the second definition show that improved observation quality may degrade system performance. We then compare and contrast these two definitions in Sect. 3.4.

Section 4 considers the POMG. Two descriptions of the value functions useful for later results are presented in Sect. 4.2, following the definition of the POMG in Sect. 4.1. In Sect. 4.3, we present a partition of the set of all observation matrices that describe the leader's observations of the follower's underlying state. Each element of this partition contains observation matrices that share a common follower policy. Section 4.4 studies the impact on the leader's value function of changing observation quality within a partition element. We (1) extend the POMDP results to the POMG under these circumstances, (2) show that the value function is Lipschitz continuous in observation quality, and then (3) present conditions that guarantee that if an observation matrix is a member of one of the partition elements and a second observation is sufficiently close to the first observation matrix, then the second observation matrix is also a member of the same partition element. In Sect. 4.5, we examine the implications of changing the observation matrix to an observation matrix in a different partition element. We show by example that crossing partition element boundaries may produce discontinuities in the leader's value function. We give conditions that guarantee that a discontinuity will be favorable to leader performance when such a discontinuity occurs. We show that when both agents are collaborative and the follower initially has complete observability of the leader, then discontinuities cannot occur. However, we show by example that improving observation quality does not necessarily improve the leader's value function, whether or not the POMG is a collaborative game.

2 Literature review

The POMDP is a sequential decision making model involving a single decision maker. Compared to the completely observed Markov decision process (i.e., the MDP; see Puterman 1994), the POMDP takes into consideration inaccurate, incomplete, and/or costly observations of the state of the system under control. However, the POMDP also introduces significant computational challenges. In the past decades, structural properties of the value function and computational procedures for the POMDP have been investigated and can be found in Smallwood and Sondik (1973), Sondik (1978), Monahan (1982), White and Scherer (1989), Lovejoy (1991), White (1991), Littman (1994), Cassandra et al. (1994), Cassandra et al. (1997), Lin et al. (1998), Lin et al. (2004), and Zhang (2010). Finite memory controllers for the POMDP have also been examined by Platzman (1977, 1980), White and Scherer (1994), Meuleau et al. (1999), and Poupart and Boutilier (2004). The finite memory controller for the POMDP in White and Scherer (1994) is extended to the POMG in this paper.

The value of information for the POMDP has been addressed in White and Harrington (1980), Zhang (2010) and Ortiz et al. (2013). This research has shown that for some sub-optimal policies, the value function may not necessarily improve if provided with more accurate state observations. We remark that there are results in the decision analysis literature (e.g., Wakker 1988) that address situations where the value of information may be negative. A comparison of the results in the POMDP literature with the condition in Wakker (1988) is a topic of future research. The definitions of observation quality presented in the POMDP literature are compared and used in this paper in order to study the value of information for POMG.

The stochastic game (Shapley 1953) is a dynamic game that is played in a sequence of stages. The state of the system evolves probabilistically over time and is controlled by one or more players. The partially observed stochastic game (POSG) is a new and relatively unexamined generalization of a stochastic game that represents a multi-agent sequential decision making problem, where the states of the game are not precisely observed by the agents and all agents make a sequence of decisions based on these partial observations. Not surprisingly, the increased modeling realism of the POSG has resulted in increased computational challenges (see Bernstein et al. 2002; Rabinovich et al. 2003). A dynamic program presented in Hansen et al. (2004) and Kumar and Zilberstein (2009) was used to prune very weakly dominated policies for both agents. Chang et al. (2014) developed a solution procedure that can generate a set of non-dominated policies for a partially observed multi-objective Markov game, from which one of two agents (the leader) can select a most preferred policy to control a dynamic system that is also affected by the control decisions of the other agent (the follower).

The value of information is a topic of considerable interest in the game theory and decision analysis literature. Much of this interest has focused on determining the value of information in games having a single decision epoch and has shown that the reward can be either improved or degraded when the decision makers are given more accurate information. For example, research by Li (2002), Chu and Lee (2006) and Leng and Parlar (2009) have examined the positive value of revealing private information to coordinate among players in the context of supply chain management problems. Bier et al. (2007) also examined the advantage of revealing private information to an attacker in a homeland security context. Bassan et al. (2003) identified a class of games where the value of information is positive. Lehrer and Rosenberg (2006) and Meyer et al. (2010) showed that the value of information is always positive in a two-person zero-sum game. On the contrary, Kamien et al. (1990) presented an example of a card game in extensive form that has a negative value of information. Zhuang and Bier (2010) argued that it is better not to reveal private information in a homeland security resource allocation problem. In multiple period games, Lehrer and Rosenberg (2010) showed that the value of information is positive for zero-sum repeated games. Zhuang et al. (2010) illustrated the use of secrecy and deception in a multiple-period, attacker-defender signaling game. To the best of our knowledge, this paper is the first to analyze the value of information of general-sum partially observed stochastic games. We study the value of improving state observation quality within the POMG framework presented in Chang et al. (2014).

3 The single agent case

3.1 POMDP problem statement

Let $\{s(t), t = 0, 1, \}$, $\{z(t), t = 1, 2, \}$, and $\{a(t), t = 0, 1, \}$ be the state, observation, and action processes, each having finite state spaces S, Z, and A, respectively. The conditional probability $p_{ij}(z,a) = P[z(t+1) = z, s(t+1) = j|s(t) = i, a(t) = a]$ is assumed given. Let $P(z,a)$ be the sub-stochastic matrix $\{p_{ij}(z,a)\}$. Note that $P[z(t+1) = z, s(t+1) = j|s(t) = i, a(t) = a] = P[z(t+1) = z|s(t+1) = j, s(t) = i, a(t) = a] \times P[s(t+1) = j|s(t) = i, a(t) = a]$, where $p_{ij}(a) = P[s(t+1) = j|s(t) = i, a(t) = a] = \sum_z p_{ij}(z,a)$ and $q_{ijz}(a) = P[z(t+1) = z|s(t+1) = j, s(t) = i, a(t) = a] = \frac{p_{ij}(z,a)}{\sum_z p_{ij}(z,a)}$ (assuming $\sum_z p_{ij}(z,a) \neq 0$) are referred to as the transition and observation probabilities, respectively. Throughout, we will assume $q_{ijz}(a)$ is independent of i and hence $q_{jz}(a) = P[z(t+1) = z|s(t+1) = j, a(t) = a]$. Let $Q(a) = \{q_{jz}(a)\}$, be the observation matrix, which we will use as our model of state observation quality.

We consider two cases for selecting actions, the perfect memory case and the finite memory case. In both cases, decision epochs are countable, $t = 0, 1, \ldots$. Let $x(0) = \{x_i(0)\}$, where $x_i(0) = P[s(0) = i]$, $d(t) = \{z(t), \ldots, z(1), a(t-1), \ldots, a(0)\}$, and $d(t,\tau) = \{z(t), \ldots, z(t-\tau+1), a(t-1), \ldots, a(t-\tau)\}$, where τ is fixed and finite and can be thought of as a design parameter. For the perfect memory case, $a(0)$ is selected on the basis of $x(0)$ and for $t \geq 1$, $a(t)$ is selected on the basis of $\{d(t), x(0)\}$. For the finite memory case, $a(t)$ is selected on the basis of $d(t,\tau)$, where $d(0,\tau)$ is assumed given.

Let $r(i,a)$ be the reward received at epoch t, given $s(t) = i$ and $a(t) = a$. The criterion we consider for the POMDP is the infinite horizon, expected total discounted reward $E\{\sum_t \beta^t r(s(t), a(t))|x(0)\}$, where $E\{.|x(0)\}$ is the expectation operator conditioned on $x(0)$ and where we will assume the discount factor β is such that $0 \leq \beta < 1$. The POMDP has two fundamental objectives. First, determine the value of the criterion for any given policy. Second, determine a policy (an optimal policy) that maximizes the criterion and its criterion value. Our interest in the POMDP formulation is to determine the value of the criterion for a given policy, to understand how this value changes as the observation matrix changes, and to extend these results to the POMG where appropriate.

3.2 Perfect memory case

We now assume all policies are perfect memory policies.

3.2.1 Value determination

Let $v^{\pi Q}(d(t))$ be the value of the criterion, assuming $\{d(t), x(0)\}$ (for notational simplicity we delete explicit dependence on x(0)), π is the policy under consideration, and Q is the observation matrix. Then $v^{\pi Q}(d(t))$ satisfies the equation

$$v^{\pi Q}(d(t)) = \sum_i r(i,a(t))P[s(t) = i|d(t)] + \beta \sum_z P[z(t+1) = z|d(t), a(t)]v^{\pi Q}(z, d(t), a(t)), \tag{1}$$

where $a(t) = \pi(d(t))$. A simple contraction mapping argument guarantees that Eq. 1 has a unique solution.

Let $v^{*Q}(d(t)) = \max_\pi v^{\pi Q}(d(t))$. Results in Bertsekas (1976) show that $v^{*Q}(d(t))$ satisfies the following optimality equation

$$\begin{aligned} v^{*Q}(d(t)) = &\max_a\{\sum_i r(i,a)P[s(t) = i|d(t)] \\ &+ \beta \sum_z P[z(t+1) = z|d(t), a]v^{*Q}(z, d(t), a)\}. \end{aligned} \tag{2}$$

Further, the perfect memory policy that causes the maximum to be attained is an optimal perfect memory policy.

3.2.2 Sufficient statistic

Due to computational tractability concerns, we seek a t-invariant sufficient statistic, observing that $d(t)$ gets large as t gets large. According to Bertsekas (1976), there exists an optimal policy that depends on $d(t)$ only through $x(t) = \{x_i(t)\}$, where $x_i(t) = P[s(t) = i|d(t)]$; i.e., $\{x(t), t = 0, 1, \ldots\}$ is a sufficient statistic for the optimization problem. Furthermore, v^{*Q} depends on $d(t)$ only through $x(t)$, and hence Eq. 2 can be transformed into

$$v^{*Q}(x) = \max_a \{xr(a) + \beta \sum_z \sigma(z, x, a) v^{*Q}(\lambda(z, x, a))\}, \tag{3}$$

where the policy that causes the maximum to be attained is an optimal policy, $xr(a) = \sum_i x_i r(i, a)$, $\sigma(z, x, a) = \sum_i x_i \sum_j p_{ij}(a) q_{jz}(a) = xP(z, a)1$ ($y1 = \sum_i y_i$ for any vector y), and $\lambda(z, x, a) = \{\lambda_j(z, x, a)\} = \frac{xP(z,a)}{\sigma(z,x,a)}$, where $\lambda_j(z, x, a) = \frac{\sum_i x_i p_{ij}(a) q_{jz}(a)}{\sigma(z,x,a)}$ when $\sigma(z, x, a) \neq 0$. We remark that $\sigma(z, x, a) = P[z(t+1) = z|d(t), a(t) = a]$ when $x(t) = x$ and $x(t+1) = \lambda(z, x, a)$ when $z(t+1) = z$, $x(t) = x$, and $a(t) = a$. Hence, $\lambda(z, x, a)$ is a form of Bayes' Rule. A result that is often exploited computationally is that the successive approximations operation preserves concavity and piecewise linearity; i.e., if v is concave and piecewise linear, then $\max_a\{xr(a) + \beta \sum_z \sigma(z, x, a) v(\lambda(z, x, a))\}$ is also concave and piecewise linear. In the limit, concavity is preserved, and hence it is also true that $v^{*Q}(x)$ is concave.

If we restrict our attention to the class of perfect memory policies that depend on $d(t)$ only through $x(t)$, then it is straightforward to show that $v^{\pi Q}(d(t))$ depends on $d(t)$ only through $x(t)$ and satisfies the equation

$$v^{\pi Q}(x) = xr(\pi(x)) + \beta \sum_z \sigma(z, x, \pi(x)) v^{\pi Q}(\lambda(z, x, \pi(x))). \tag{4}$$

3.2.3 State observation quality

We now consider the first definition of observation quality presented in White and Harrington (1980) and Zhang (2010). Proofs of all results are presented in Appendix 1.

Definition 1 Observation matrix Q' is at least as informative as observation matrix Q if there exists a stochastic matrix R such that $Q'(a)R(a) = Q(a)$ for all a.

Theorem 1 *Assume that observation matrix Q' is at least as informative (in terms of Definition 1) as observation matrix Q. Let $v^{\pi Q}(x)$ be the solution of Eq. 4, and assume $v^{\pi Q}(x)$ is concave in x. Then, $v^{\pi Q}(x) \leq v^{\pi Q'}(x)$ for all x.*

A policy is said to be Q-adaptive if when given a more informative observation matrix, the policy gives in return improved performance. Theorem 1 states that if $v^{\pi Q}$ is concave, then π is Q-adaptive. Theorem 1 leads to the following results.

Corollary 1 (*a*) *Assume π is optimal for observation matrix Q, π' is optimal for observation matrix Q', and that observation matrix Q' is at least as informative as observation matrix Q (in terms of Definition 1). Then, $v^{\pi Q}(x) \leq v^{\pi' Q'}(x)$ for all x.* (*b*) *Assume Q has rank 1 and*

$Q'' = I$, where I is the identity matrix, and let π and π'' be optimal policies for observation matrices Q and Q'', respectively. Let Q' be any observation matrix, and assume π' is an optimal policy for Q'. Then, $v^{\pi Q}(x) \leq v^{\pi' Q'}(x) \leq v^{\pi'' Q''}(x)$ for all x.

Corollary 1(a) results from the fact that an optimal policy always produces a concave value function. Corollary 1(b) notes that there is an observation matrix (or family of matrices) for which any given observation matrix is at least as informative, and there is an observation matrix (or family of matrices) that is at least as informative as any given observation matrix. The POMDP based on the former observation matrix (having rank 1) is called the completely unobserved case, and the POMDP based on the latter observation matrix (the identity matrix) is called the completely observed case (or simply, the MDP). The value functions of these special cases represent lower and upper bounds, respectively, on the value function of the general case.

3.3 Finite memory case

We now state an alternative definition of observation quality presented in Ortiz et al. (2013), assuming that the underlying state of the system is close to completely observed (e.g., true for many inventory systems) and that all policies considered are finite-memory.

Definition 2 Let $Q(\epsilon) = I(1-\epsilon) + P\epsilon$, P is a stochastic matrix with zeros on the diagonal, and $\epsilon \geq 0$ is small. Observation matrix $Q(\epsilon')$ is at least as informative as observation matrix $Q(\epsilon)$ if and only if $\epsilon' \leq \epsilon$.

Results in Ortiz et al. (2013) state that for finite-memory policy π,

$$v^{\pi Q}(x) = \lambda^{\tau}(d(t,\tau), x')\gamma(d(t,\tau)),$$

where $x = x(t) = \lambda^{\tau}(d(t,\tau), x')$ given $x' = x(t-\tau)$ and $\gamma(d(t,\tau))$ is polynomial in ϵ; i.e., there is a sequence of (easily computed) vectors $\{\alpha^k(d(t,\tau))\}$ such that $\gamma(d(t,\tau)) = \sum_{k=0}^{\infty}\epsilon^k\alpha^k(d(t,\tau))$. Thus, for sufficiently small $\epsilon > 0$, $\sum_{k=0}^{\infty}\epsilon^k\alpha^k(d(t,\tau))$ is well-defined and can be approximated by $\alpha^0(d(t,\tau)) + \epsilon\alpha^1(d(t,\tau))$, and hence the signs of the scalar elements of the vector $\alpha^k(d(t,\tau))$ for $k = 1$ determine whether or not $v^{\pi Q}(x)$ will increase or decrease as a function of ϵ. Not unexpectedly, if the finite-memory policy under consideration achieves the maximum for the completely observed MDP, then it is shown in Ortiz et al. (2013) that the signs of all elements of the vector $\alpha^1(d(t,\tau))$ are negative; hence, this policy improves system performance if given improved observation quality.

3.4 A comparison of definitions of observation quality

We now show that if there exists a stochastic matrix R such that $Q(\epsilon)R = Q(\epsilon')$, then for sufficiently small ϵ and ϵ', $\epsilon' \geq \epsilon > 0$. Further, we show that for sufficiently small ϵ and ϵ', $\epsilon' \geq \epsilon > 0$, there may not exist a stochastic matrix R such that $Q(\epsilon)R = Q(\epsilon')$. Thus, for an observation matrix having the specialized form $Q(\epsilon) = I(1-\epsilon) + P\epsilon$, where $\epsilon \geq 0$ is small and P is a stochastic matrix with zeros on the diagonal, the definition "$Q(\epsilon)$ is at least as informative as $Q(\epsilon')$ when $\epsilon' \geq \epsilon > 0$" is more general than the definition "$Q(\epsilon)$ is at least as informative as $Q(\epsilon')$ when there exists a stochastic matrix R such that $Q(\epsilon)R = Q(\epsilon')$".

We now determine that the existence of a stochastic matrix R such that if $Q(\epsilon)R = Q(\epsilon')$ implies $\epsilon' \geq \epsilon > 0$ for sufficiently small ϵ and ϵ'. Equivalently, if $\epsilon' < \epsilon$, then $Q(\epsilon)^{-1}Q(\epsilon')$ cannot be stochastic, assuming the existence of $Q(\epsilon)^{-1}$. Assume $\epsilon' < \epsilon$, and let $\kappa = \frac{\epsilon}{1-\epsilon}$ and $\kappa' = \frac{\epsilon'}{1-\epsilon'}$. Then, $Q(\epsilon)^{-1}Q(\epsilon') = \frac{(1-\epsilon')}{1-\epsilon}(I + \kappa P)^{-1}(I + \kappa' P)$. Since $\kappa < 1$ for ϵ

sufficiently small and $(I+\kappa P)^{-1} = I - \kappa P(I+\kappa P)^{-1}$, it follows that $(I+\kappa P)^{-1} = \sum(-1)^n \kappa^n P^n$, where the sum is over all $n = 0, 1, 2, \ldots$. It is then straightforward to show that $(I+\kappa P)^{-1}(I+\kappa' P) = I + (\kappa'-\kappa)P(I - \kappa P + \kappa^2 P^2 - \ldots)$, which can be approximated by $I + (\kappa' - \kappa)P$ for small κ. Thus, for $i \neq j$, the $(i,j)^{th}$ term of $R = Q(\epsilon)^{-1}Q(\epsilon')$ is $r_{ij} = (1-\epsilon')(\kappa'-\kappa)p_{ij}/(1-\epsilon)$, which is negative since $\epsilon' < \epsilon$, and the result is proved.

If $\epsilon' > \epsilon > 0$ (both small), does there exist a stochastic matrix R such that $Q(\epsilon)R = Q(\epsilon')$? Such a stochastic matrix exists when the state and observation spaces have cardinality 2 (let $r_{11} = r_{22} = (1-\epsilon-\epsilon')/(1-2\epsilon)$ and $r_{12} = r_{21} = (\epsilon'-\epsilon)/(1-2\epsilon)$, for $\epsilon < \frac{1}{2}$ and $(\epsilon'+\epsilon) < 1$). However, the existence of such a stochastic matrix R is not guaranteed for larger dimensional problems. For example, let

$$Q(\epsilon) = \begin{bmatrix} 1-\epsilon & \epsilon & 0 \\ 0 & 1-\epsilon & \epsilon \\ 0 & \epsilon & 1-\epsilon \end{bmatrix}.$$

Then,

$$Q(\epsilon)^{-1} = \frac{1}{(1-\epsilon)(1-2\epsilon)} \begin{bmatrix} 1-2\epsilon & -\epsilon(1-\epsilon) & \epsilon^2 \\ 0 & (1-\epsilon)^2 & -\epsilon(1-\epsilon) \\ 0 & -\epsilon(1-\epsilon) & (1-\epsilon)^2 \end{bmatrix},$$

where the inverse exists when $\epsilon < \frac{1}{2}$. We observe that the $(1,3)$ entry of $R = Q(\epsilon)^{-1}Q(\epsilon')$ is $r_{13} = -\frac{\epsilon(\epsilon'-\epsilon)}{(1-\epsilon)(1-2\epsilon)} < 0$, and hence R, although unique, is not stochastic. Thus, if $\epsilon' > \epsilon > 0$, it is not guaranteed that there exists a stochastic matrix R such that $Q(\epsilon)R = Q(\epsilon')$.

4 Partially observed Markov game

4.1 Problem statement

Thus far, we have explored the impact of changes to the observation quality of the underlying state process on system performance for the POMDP and have presented results that address the question: will improved observation quality improve system performance? Given this context, we now investigate this question for the infinite horizon, expected total discounted POMG.

We assume that the POMG has two agents. The first agent, the leader, chooses its policy. Then the second agent, the follower, selects its policy with complete knowledge of the policy selected by the leader. Let $\{s^k(t), t = 0, 1, \ldots\}$, $\{z^k(t), t = 1, \ldots\}$, and $\{a^k(t), t = 0, 1, \ldots\}$ be the state, observation, and action processes for agent $k \in \{L = \text{Leader}, F = \text{Follower}\}$, each having finite state spaces S^k, Z^k, and A^k, respectively. Let $s(t) = \{s^L(t), s^F(t)\}$, $z(t) = \{z^L(t), z^F(t)\}$, and $a(t) = \{a^L(t), a^F(t)\}$, where $z^k(t)$ is the observation received by agent k of the other agent's state. The conditional probability $p_{ij}(z,a) = P[z(t+1) = z, s(t+1) = j | s(t) = i, a(t) = a]$ is assumed given. Let $P(z,a)$ be the sub-stochastic matrix $\{p_{ij}(z,a)\}$.

Let the information pattern at time t of finite length τ for agent k be $d^k(t,\tau) = \{s^k(t), ..., s^k(t-\tau+1), z^k(t), ..., z^k(t-\tau+1), a^k(t-1), ..., a^k(t-\tau)\}$, hence, $d^k(t,\tau) = \{s^k(t), z^k(t), a^k(t-1), d^k(t-1, \tau-1)\}$. And let $y^k(t) = \{P(d^l(t,\tau)|d^k(t))\}$ for $l \neq k$, where $y^k(t)$ is a "belief" array that indicates what agent k can infer about the other agent's information pattern, i.e., $d^l(t,\tau), l \neq k, l, k \in \{L, F\}$. Denote $d^k(t) = \{z^k(t), ..., z^k(1), s^k(t), ..., s^k(0), a^k(t-1), ..., a^k(0), y^k(0)\}$ when $t \geq 1$, where $y^k(0) = \{P(d^l(0,\tau))\}$, hence, $d^k(t) = \{z^k(t), s^k(t), a^k(t-1), d^k(t-1)\}$. The decision epochs are $t = 0, 1, \ldots$, and agent k selects $a^k(t)$ on the basis of information pattern $d^k(t, \tau)$. Hence, when selecting an action, we assume that agent k knows the current and τ most recent observations of the other agent's state, its current and τ most recent state values, and the τ most recent actions it has selected. Let $v^k(\pi^L, \pi^F, Q)(d^k(0))$ be the value of agent k's criterion, assuming the leader and follower policies are π^L and π^F, respectively, and $Q = \{P(z^L|s^F)\}$ is the leader's observation matrix. We let $d(t, \tau) = \{d^L(t, \tau), d^F(t, \tau)\}$. It will be convenient to describe the policy pair (π^L, π^F) as $\{P(a(t)|d(t, \tau))\}$, where $P(a(t)|d(t, \tau)) = P(a^L(t)|d^L(t, \tau))P(a^F(t)|d^F(t, \tau))$.

The criterion we consider for agent k is the infinite horizon, expected total discounted reward; i.e., $v^k(\pi^L, \pi^F, Q)(d^k(0)) = E\{\sum_t \beta^t r^k(s(t), a(t))|d^k(0)\}$, where $E\{.|d^k(0)\}$ is the expectation operator conditioned on $d^k(0)$, $\beta \in [0, 1)$ is the discount factor, and $r^k(i, a)$ is the scalar reward received by agent k at epoch t, given $s(t) = i$ and $a(t) = a$.

Let $\mathcal{Q}$ be the set of all stochastic matrices and hence the set of all observation matrices $Q = \{P(z^L|s^F)\}$. We remark that $\mathcal{Q}$ is equivalent to the set of all elements in $R^{|S^F|} \times R^{(|Z^L|-1)}$ such that $q_{jz} \geq 0$ for all $j \in S^F$ and $z \in Z^L$, and $\sum_{z=1}^{|z^L|-1} q_{jz} \leq 1$ for all $j \in S^F$. For example, if $|S^F| = |Z^L| = 2$, then $\mathcal{Q}$ is equivalent to $\{(x, y) : 0 \leq x \leq 1, 0 \leq y \leq 1\}$. Hence, $\mathcal{Q}$ is compact.

Assume initial conditions $d^L(0)$ and $d^F(0)$ are given. Let Π^L and Π^F be the set of all policies from which the leader and the follower can choose, respectively. Both Π^L and Π^F are assumed to contain only finite-memory policies. Let the response function of the follower $\pi^* : \Pi^L \times \mathcal{Q} \to \Pi^F$ be such that $\forall \pi^L \in \Pi^L$,

$$v^F(\pi^L, \pi^*(\pi^L, Q))(d^F(0)) \geq v^F(\pi^L, \rho^F)(d^F(0)), \forall \rho^F \in \Pi^F.$$

Then, the equilibrium conditions for the POMG are:

(1) $v^L(\pi^L, \pi^*(\pi^L, Q))(d^L(0)) \geq v^L(\rho^L, \pi^*(\rho^L, Q)))(d^L(0)), \forall \rho^L \in \Pi^L$;
(2) $\forall \rho^L \in \Pi^L, v^F(\rho^L, \pi^*(\rho^L, Q))(d^F(0)) \geq v^F(\rho^L, \rho^F)(d^F(0)), \forall \rho^F \in \Pi^F$.

Hence, neither the leader nor the follower can improve its performance by deviating from the equilibrium condition.

Our focus in this paper is to determine $v^L(\rho^L, \pi^*(\rho^L, Q), Q)(d^L(0))$ for any given $\rho^L \in \Pi^L$. We assume that determining this scalar value is a critical step in determining the most preferred leader policy in Π^L. We note that a genetic algorithm was used in Chang et al. (2014) to determine the most preferred leader policy, using $v^L(\rho^L, \pi^*(\rho^L, Q), Q)(d^L(0))$ as the fitness measure.

4.2 Descriptions of v^k, $k \in \{L, F\}$

We now present two descriptions of v^k in Proposition 1 and Proposition 2 that will be useful in determining results below. The first description describes v^k in a manner analogous to the description of the optimality equation for the POMDP. The second description takes advantage of the fact that both leader and follower policies are assumed to be finite memory policies.

Proposition 2 in Chang et al. (2014) implies that $\{d^k(t, \tau), y^k(t)\}$ is a sufficient statistic for $\{d^k(t)\}$, and hence $v^k(\pi^L, \pi^F, Q)(d^k(t)) = v^k(\pi^L, \pi^F, Q)(d^k(t, \tau), y^k(t))$.

Let one-period information for agent k be $\varsigma^k(t) = \{z^k(t), s^k(t), a^k(t-1)\}$ and $\varsigma(t) = \{\varsigma^L(t), \varsigma^F(t)\}$. Define

(1) $\sigma^k(\varsigma^k(t+1), d^k(t, \tau), y^k(t)) = P(\varsigma^k(t+1)|d^k(t)) = \sum_{\varsigma^l(t+1)} \sum_{d^l(t,\tau)} P(z(t+1), s(t+1)|s(t), a(t))P(a(t)|d(t, \tau))P(d^l(t, \tau)|d^k(t, \tau)), l \neq k$

(2) $\lambda^k(\varsigma^k(t+1), d^k(t,\tau), y^k(t))$ is the stochastic array with scalar element

$$P(\varsigma^l(t+1), d^l(t,\tau-1)|\varsigma^k(t+1), d^k(t)) = \frac{P(\varsigma(t+1), d^l(t,\tau-1)|d^k(t))}{P(\varsigma^k(t+1)|d^k(t))},$$

where

$$\begin{aligned}&P(\varsigma(t+1), d^l(t,\tau-1)|d^k(t))\\ &\quad= \sum_{\varsigma^l(t-\tau+1)} P(z(t+1), s(t+1)|s(t), a(t))P(a(t)|d(t,\tau))P(d^l(t,\tau)|d^k(t)), l \neq k,\end{aligned}$$

and where we assume $P(\varsigma^k(t+1)|d^k(t)) \neq 0$.

Proposition 1 *For policies π^L and π^F and observation matrix Q,*

$$\begin{aligned}&v^k(\pi^L, \pi^F, Q)(d^k(t)) = v^k(\pi^L, \pi^F, Q)(d^k(t,\tau), y^k(t))\\ &\quad= \sum_{s(t)}\sum_{a(t)} r^k(s(t), a(t)) \sum_{d^l(t,\tau)} P(a(t)|d(t,\tau))P(d^l(t,\tau)|d^k(t))\\ &\qquad+ \beta \sum_{\varsigma^k(t+1)} \sigma^k(\varsigma^k(t+1), d^k(t,\tau), y^k(t))\\ &\qquad\times v^k(\pi^L, \pi^F, Q)(\{\varsigma^k(t+1), d^k(t,\tau-1)\}, \lambda^k(\varsigma^k(t+1), d^k(t,\tau), y^k(t))), l \neq k.\end{aligned}$$

We now present a sufficient statistic and a structured result for v^k after the following definition. Let g^k be the solution to the equation

$$\begin{aligned}&g^k(d(t,\tau), \pi^L, \pi^F, Q)\\ &\quad= R^k(d(t,\tau), \pi^L, \pi^F) + \beta \sum_{\varsigma(t+1)} P(\varsigma(t+1)|d(t,\tau), \pi^L, \pi^F, Q)\\ &\qquad\times g^k(\{\varsigma(t+1), d(t,\tau-1)\}, \pi^L, \pi^F, Q),\end{aligned}$$

where

$$R^k(d(t,\tau), \pi^L, \pi^F) = \sum_{a(t)} r^k(s(t), a(t))P(a(t)|d(t,\tau)).$$

Proposition 2 *For policies π^L and π^F and observation matrix Q,*

$$\begin{aligned}&v^k(\pi^L, \pi^F, Q)(d^k(t)) = v^k(\pi^L, \pi^F, Q)(d^k(t,\tau), y^k(t))\\ &\quad= \sum_{d^l(t,\tau)} P(d^l(t,\tau)|d^k(t))g^k(d(t,\tau), \pi^L, \pi^F, Q), l \neq k.\end{aligned}$$

4.3 Partition of observation matrices

Let $K(\pi^L, \pi^F) = \{Q \in \mathscr{Q} : v^F(\pi^L, \pi^F, Q)(d^F(0)) \geq v^F(\pi^L, \rho^F, Q)(d^F(0)), \forall \rho^F \in \Pi^F\}$. Thus, $K(\pi^L, \pi^F)$ is the set of all matrices $Q = \{P(z^L|s^F)\}$ such that given π^L, the follower will select policy π^F (i.e., $\pi^*(\pi^L, Q) = \pi^F$). We assume that the policies in Π^F have been selected so that if $\pi^F, \rho^F \in \Pi^F$ and $\pi^F \neq \rho^F$, then $K(\pi^L, \pi^F) \cap K(\pi^L, \rho^F)$ may be non-empty but has (Lebesgue) measure zero. Thus, $\{K(\pi^L, \rho^F) : \rho^F \in \Pi^F\}$ is a finite partition of $\mathscr{Q}$ (permitting non-empty intersections when equalities occur); hence, there exists at least one element of this partition that has a non-empty interior. We assume

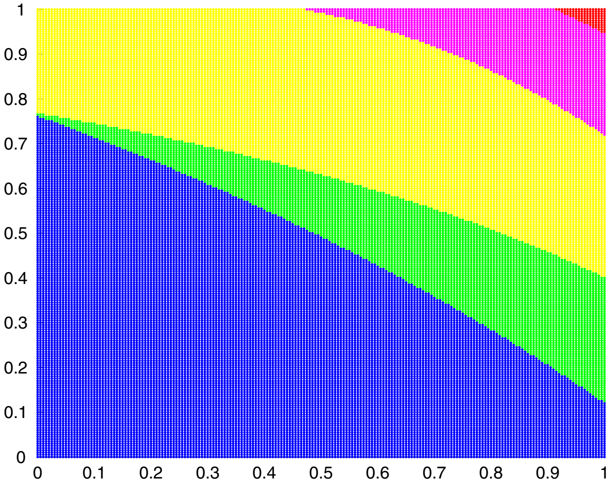

Fig. 1 An example of a partition over $\mathscr{Q} = \{(x, y) : 0 \leq x \leq 1, 0 \leq y \leq 1\}$

that the follower selects π^F in order to maximize its criterion value. Thus, if Q is a member of only $K(\pi^L, \pi^F)$, then the follower will select π^F in response to the leader selecting π^L. If $Q \in K(\pi^L, \pi^F) \cap K(\pi^L, \rho^F)$, then either π^F or ρ^F may be selected. We assume that the leader has complete knowledge of $\{K(\pi^L, \rho^F) : \rho^F \in \Pi^F\}$. We remark that $K(\pi^L, \pi^F)$ for each $\pi^F \in \Pi^F$ is compact. Figure 1 is an illustration of a partition over $\mathscr{Q} = \{(x, y) : 0 \leq x \leq 1, 0 \leq y \leq 1\}$, where different colored regions correspond to different response policies ρ^F.

4.4 Changing Q within a partition element

We now examine the impact on $v^L(\pi^L, \pi^F, Q)(d^L(t))$ of changing Q to Q' when $Q, Q' \in K(\pi^L, \pi^F)$. We begin by extending Theorem 1 and Corollary 1 to the POMG. We then show that $v^L(\pi^L, \pi^F, Q)(d^L(t))$ is Lipschitz continuous on $K(\pi^L, \pi^F)$. Finally, we present conditions that guarantee two sufficiently close observation matrices are members of the same element of the partition $\{K(\pi^L, \rho^F) : \rho^F \in \Pi^F\}$.

Corollary 2 *Assume $Q, Q' \in K(\pi^L, \pi^F)$ and there is a $R \in \mathscr{Q}$ such that $Q'R = Q$.*

(1) If $v^L(\pi^L, \pi^F, Q)(d^L(t, \tau), y^L(t))$ is concave in $y^L(t)$ for $d^L(t, \tau)$, then

$$v^L(\pi^L, \pi^F, Q')(d^L(t, \tau), y^L(t)) \geq v^L(\pi^L, \pi^F, Q)(d^L(t, \tau), y^L(t)).$$

(2) Let π^L be such that

$$v^L(\pi^L, \pi^F, Q)(d^L(t, \tau), y^L(t)) \geq v^L(\rho^L, \pi^*(\rho^L, Q), Q)(d^L(t, \tau), y^L(t))$$

for all $\rho^L \in \Pi^L$. Then,

$$v^L(\pi^L, \pi^F, Q')(d^L(t, \tau), y^L(t)) \geq v^L(\pi^L, \pi^F, Q)(d^L(t, \tau), y^L(t)).$$

Thus, within an element of $\{K(\pi^L, \rho^F) : \rho^F \in \Pi^F\}$, Corollary 2 gives conditions that insure improved observation quality (based on Definition 1) will improve the leader's value function. We remark that Corollary 2 essentially extends the notion of Q-adaptivity to a pair of policies (π^L, π^F) when $Q, Q' \in K(\pi^L, \pi^F)$.

We note that Π^L and Π^F may not contain a π^L and a $\pi^*(\pi^L, Q)$ such that these equilibrium conditions hold for all initial conditions. Determining conditions that guarantee the existence of such policies is a future research topic.

We now show that for all $d^L(t)$ and for any pair of finite-memory policies (π^L, π^F), $v^L(\pi^L, \pi^F, Q)(d^L(t))$ is Lipschitz continuous in Q on $K(\pi^L, \pi^F)$.

Proposition 3 *For all $(d^L(t, \tau), y^L(t))$ and for any pair of finite-memory policies (π^L, π^F), $v^L(\pi^L, \pi^F, Q)(d^L(t, \tau), y^L(t))$ is Lipschitz continuous in Q on $K(\pi^L, \pi^F)$.*

Thus, as long as we remain in one of the elements of the partition $\{K(\pi^L, \rho^F) : \rho^F \in \Pi^F\}$, $v^L(\pi^L, \pi^F, Q)(d^L(t, \tau), y^L(t))$ is Lipschitz continuous in Q.

We now present conditions that guarantee that two observation matrices are members of the same partition element. Consider the following definitions:

(1) for any vector g, $||g|| = \max_s |g(s)|$;
(2) for any matrix Q, $||Q|| = \max_j \sum_{z \in Z} |q_{jz}|$.

Define $B(\rho^F) = v^F(\pi^L, \pi^F, Q')(d^F(0)) - v^F(\pi^L, \rho^F, Q')(d^F(0))$ for a given $d^F(0)$, and $b = \min\{B(\rho^F) : \rho^F \in \Pi^F, \rho^F \neq \pi^F\} \geq 0$.

Proposition 4 *Assume $Q' \in K(\pi^L, \pi^F)$. If Q is such that $||Q - Q'|| \leq \frac{b(1-\beta)^2}{2\beta M}$ where $M = \max_s \max_a r^L(s, a)$, then $Q \in K(\pi^L, \pi^F)$.*

Hence, as long as observation matrices Q and Q' are close enough, they are in the same partition element which shares the same response policy. Assume current leader policy favors more accurate observation quality, Proposition 4 indicates how much the observation quality can improve safely without the follower changing its policy. Section 4.5 will show that when the observation quality is changed large enough so that the follower changes its policy, discontinuities will occur and these discontinuities can be beneficial or not beneficial to the leader.

4.5 Changing Q across partition elements

We now examine the impact of changing Q to Q' on $v^L(\pi^L, \pi^F, Q)(d^L(t, \tau), y^L(t))$ when $Q \in K(\pi^L, \pi^F)$, $Q' \in K(\pi^L, \rho^F)$, and $\pi^F \neq \rho^F$. Without loss of generality, assume $Q^* \in K(\pi^L, \pi^F) \cap K(\pi^L, \rho^F)$, $\{Q_n\}$ is a sequence in $K(\pi^L, \pi^F)$ that converges to Q^*, and $\{Q'_n\}$ is a sequence in $K(\pi^L, \rho^F)$ that converges to Q^*. Then, from the Proposition 3 and the compactness of $K(\pi^L, \pi^F)$ and $K(\pi^L, \rho^F)$,

$$\lim_{n\to\infty} v^L(\pi^L, \pi^F, Q_n)(d^L(0)) = v^L(\pi^L, \pi^F, Q^*)(d^L(0))$$
$$\lim_{n\to\infty} v^L(\pi^L, \rho^F, Q'_n)(d^L(0)) = v^L(\pi^L, \rho^F, Q^*)(d^L(0)).$$

However, there is no guarantee that $v^L(\pi^L, \pi^F, Q^*)(d^L(0)) = v^L(\pi^L, \rho^F, Q^*)(d^L(0))$, and hence, at a boundary there can be discontinuities. Figure 2 presents a 3-dimensional view of possible discontinuities at the boundaries of the partition elements over $\mathscr{Q} = \{(x, y) : 0 \leq x \leq 1, 0 \leq y \leq 1\}$, where different colored regions correspond to different response policies ρ^F.

We now show by example a variety of ways that v^L can depend on Q as Q moves across boundaries in the partition $\{K(\pi^L, \pi^F), \pi^F \in \Pi^F\}$. In order to reduce computational complexity, we assume that $a^L(t)$ depends only on $\{s^L(t), z^L(t)\}$ and that the follower can completely observe the leader's information $\{s^L(t), z^L(t)\}$. Parameter values for all examples are presented in Appendix 2.

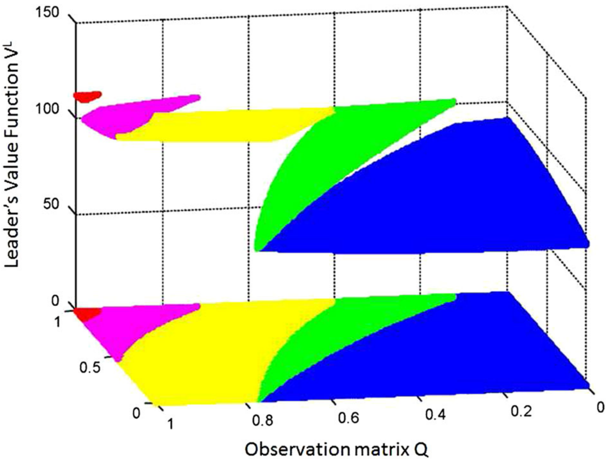

Fig. 2 An example of discontinuities at the boundaries of the partition elements over $\mathscr{Q} = \{(x, y) : 0 \leq x \leq 1, 0 \leq y \leq 1\}$

Example 1 Let $|S^L| = |A^L| = |Z^L| = |S^F| = |A^F| = 2$. Let $Q' = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ and $R(\epsilon) = \begin{bmatrix} 1-\epsilon & \epsilon \\ \epsilon & 1-\epsilon \end{bmatrix}$ where $\epsilon > 0$ and is given. Define $Q(x) = Q'R^x$, then $Q(x)$ is a stochastic matrix for all $x \geq 0$ and $Q(x_1)$ is at least as informative as $Q(x_2)$ if $0 \leq x_1 \leq x_2$.
Note, the Markov chain constructed by R is aperiodic and irreducible, hence there exists a unique Q^* such that $Q^* = \lim_{x\to\infty} Q'R^x$ and $Q^* = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}$, corresponding to the completely unobserved case. In addition, the second largest eigenvalue value of R is $1 - 2\epsilon$; hence $Q(x)$ converges to Q^* faster as ϵ increases.

Figure 3 shows the changes in the leader's value function for a given leader's policy as observation quality degrades for six examples. In these examples, only $r^L(s(t), a(t))$ was varied; $P(z(t+1), s(t+1)|s(t), a(t))$ and $r^F(s(t), a(t))$ remained unchanged. A discontinuity can occur when the change in $Q(x)$ is great enough to cause a change in the follower's policy. Note that the leader completely observes the follower when $x = 0$, whereas when x is large ($x \geq 20$ in these examples), the leader receives no information about the follower from observations. Regarding the examples in Fig. 3, all six have three discontinuities. As x increases, we note discontinuities that produce abrupt increases or decreases in the leader's value function, and between the discontinuities we note monotone increasing or decreasing value function performance.

The examples have shown that if the leader's observation matrix changes from $Q \in K(\pi^L, \pi^F)$ to $Q' \in K(\pi^L, \rho^F)$, $\pi^F \neq \rho^F$, then the leader's value function may experience an abrupt change of value due to discontinuities that can occur at partition boundaries and that these changes can be favorable or unfavorable. We now present sufficient conditions that guarantee a favorable change.

Let $N = (|Z^L||Z^F||S^L||S^F||A^L||A^F|)^\tau$, and assume μ is a one-to-one, onto mapping from $\{d(t, \tau)\}$, the set of all $d(t, \tau)$, to $\{1, 2, ..., N\}$. Thus, μ totally orders $\{d(t, \tau)\}$.

A function $f : \{d(t, \tau)\} \to R$ is said to be isotone (with respect to μ) if and only if $\mu(d(t, \tau)) \leq \mu(d'(t, \tau))$ implies $f(d(t, \tau)) \leq f(d'(t, \tau))$.

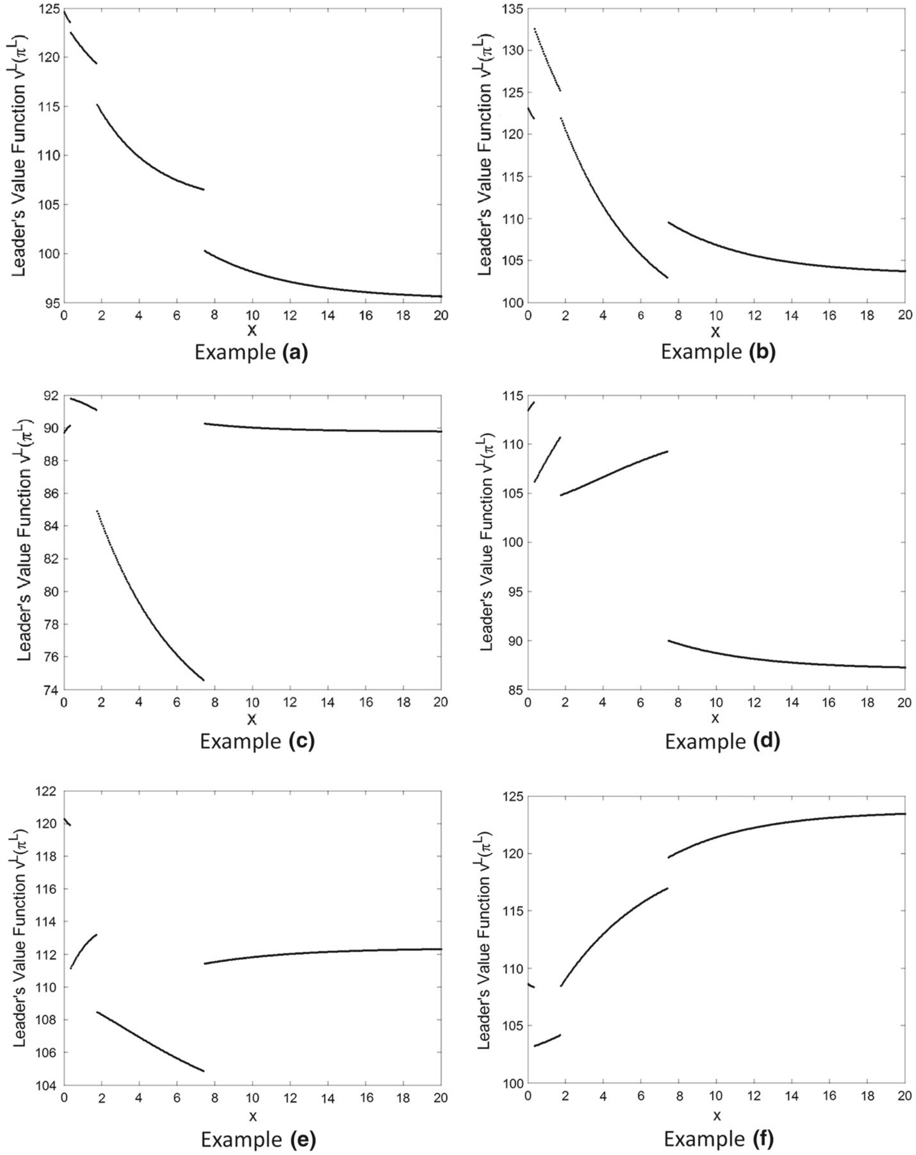

Fig. 3 A variety of changes to the leader's value function as observation quality degrades

For any $\pi' \in \Pi^F$ and $Q \in \mathscr{Q}$, let

$$q(k|d(t,\tau), Q, \pi^L, \pi') = \sum P(\varsigma(t+1)|d(t,\tau), Q, \pi^L, \pi'),$$

where the sum is over all $\varsigma(t+1)$ such that $\mu(\{\varsigma(t+1), d(t,\tau-1)\}) \geq k$.

Lemma 1 *Assume:*

(1) $R^L(d(t,\tau), \pi^L, \pi^F)$ *is isotone in* $d(t,\tau)$*,*
(2) $q(k|d(t,\tau), Q, \pi^L, \pi^F)$ *is isotone in* $d(t,\tau)$ *for all* k*,*

Then, $g^L(d(t,\tau), Q, \pi^L, \pi^F)$ *is isotone in* $d(t,\tau)$*.*

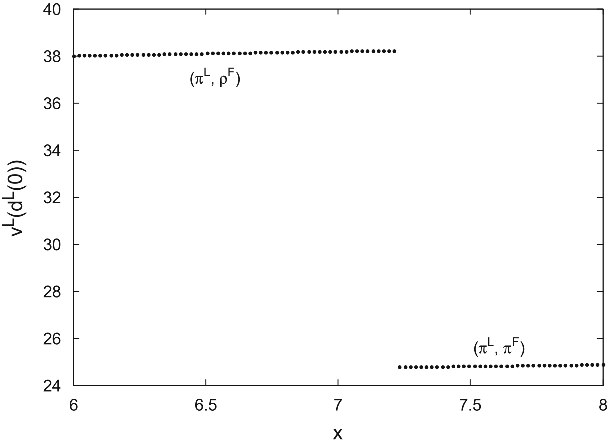

Fig. 4 Favorable change in leader's value function across the boundary

Proposition 5 *Assume:*

(1) $Q^* \in K(\pi^L, \pi^F) \cap K(\pi^L, \rho^F), \pi^F \neq \rho^F$,
(2) $R^L(d(t, \tau), \pi^L, \pi')$ *is isotone in* $d(t, \tau)$ *for* $\pi' \in \{\pi^F, \rho^F\}$,
(3) $q(k|d(t, \tau), Q^*, \pi^L, \pi')$ *is isotone in* $d(t, \tau)$ *for all* k *for* $\pi' \in \{\pi^F, \rho^F\}$,
(4) $R^L(d(t, \tau), \pi^L, \rho^F) \geq R^L(d(t, \tau), \pi^L, \pi^F)$ *for all* $d(t, \tau)$,
(5) $q(k|d(t, \tau), Q^*, \pi^L, \rho^F) \geq q(k|d(t, \tau), Q^*, \pi^L, \pi^F)$ *for all* k *and all* $d(t, \tau)$.

Then, $g^L(d(t, \tau), Q^*, \pi^L, \rho^F) \geq g^L(d(t, \tau), Q^*, \pi^L, \pi^F)$ *for all* $d(t, \tau)$*, and hence* $v^L(\pi^L, \rho^F, Q^*)(d^L(0)) \geq v^L(\pi^L, \pi^F, Q^*)(d^L(0))$.

Proposition 5 presents conditions involving both $R^L(d(t, \tau), \pi^L, \pi')$ and $P(\varsigma(t + 1)|d(t, \tau), Q^*, \pi^L, \pi')$ that suggest a change in observation quality from $Q \in K(\pi^L, \pi^F)$ to $Q' \in K(\pi^L, \rho^F)$, where Q and Q' are both close to $Q^* \in K(\pi^L, \pi^F) \cap K(\pi^L, \rho^F)$, will improve the leader's performance.

It is easily shown that the result in Proposition 5 holds if

$$\max_{d(t,\tau)} R^L(d(t, \tau), \pi^L, \pi^F) \leq \min_{d(t,\tau)} R^L(d(t, \tau), \pi^L, \rho^F), \tag{5}$$

which does not require assumptions on $P(\varsigma(t+1)|d(t, \tau), Q^*, \pi^L, \pi')$. We remark, however, that (5) is considerably stronger than Assumption (4) in Proposition 5 and hence may be more difficult to be satisfied than the assumptions in Proposition 5.

Example 2 Consider the problem in Example 1, assuming the leader's information can be only partially observed by the follower and $P(d^L(0)|d^F(0)) = 1$. Figure 4 shows that a favorable change in leader's value function can occur at partition boundaries when the assumptions in Proposition 5 are satisfied.

We now examine the situation where the two agents are collaborative (share the same objective; i.e. $r^L = r^F$) and at least the follower has complete knowledge of the leader's initial finite-memory state of knowledge ($d^L(0, \tau)$). Under these conditions there will not be a discontinuity in crossing a boundary of the partition $\{K(\pi^L, \pi^F) : \pi^F \in \Pi^F\}$.

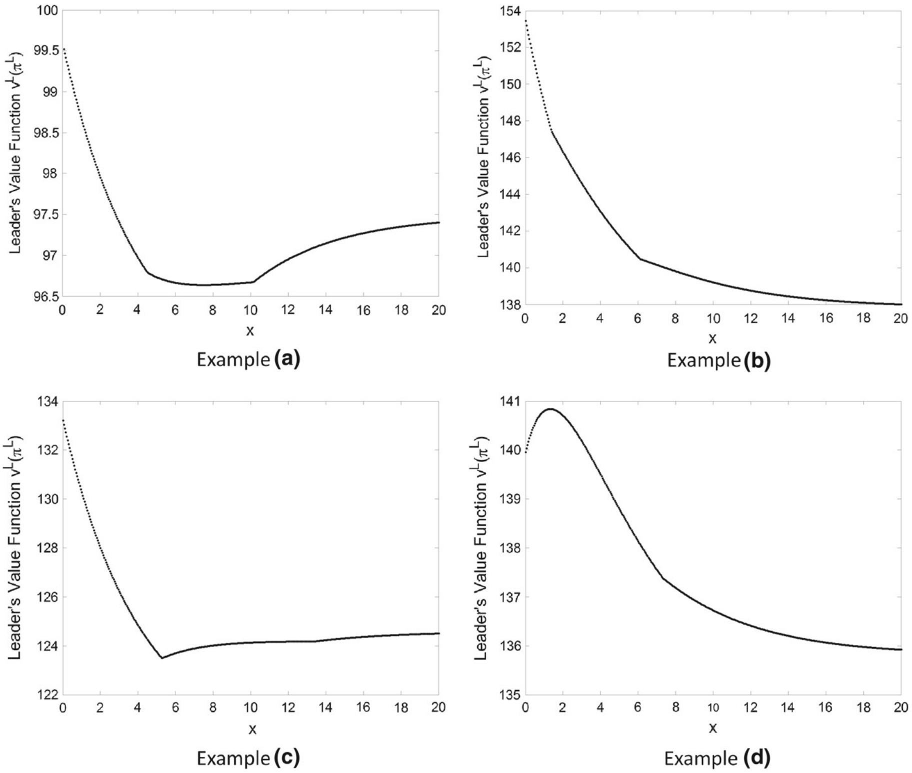

Fig. 5 Changes to the leader's value function as observation quality degrades under conditions of Proposition 6

Proposition 6 *For* $d^L(0) = \{d^L(0, \tau), y^L(0)\}$ *and* $d^F(0) = \{d^F(0, \tau), y^F(0)\}$, *assume:*

(1) $P(d^L(0, \tau)|d^F(0)) = 1$,
(2) $Q^* \in K(\pi^L, \pi^F) \cap K(\pi^L, \rho^F), \pi^F \neq \rho^F$,
(3) $r^L = r^F$,

Then, $v^L(\pi^L, \pi^F, Q^*)(d^L(0)) = v^L(\pi^L, \rho^F, Q^*)(d^L(0))$.

We now provide several illustrative examples under the conditions assumed in Proposition 6.

Example 3 Assume the problem in Example 1 under the assumptions of Proposition 6. Figure 5 shows the changes to the leader's value function as observation quality degrades for four randomly generated examples. All discontinuity points disappear, and the value function of the leader v^L is continuous with respect to observation matrix Q. However, the slope of the value function v^L and its sign can still be negative or positive due to the changes of the best response policy ρ^F.

The implications of these illustrative examples are that even if the two agents are totally collaborative (i.e., $r^L = r^F$) and initially at least the follower has complete visibility of the leader's state (i.e., $y^F(0)$ identifies $d^L(0, \tau)$ with probability 1), improved observation quality for the leader may or may not improve system performance, before or after a boundary is crossed. These results indicate that greater visibility between even collaborative agents may not result in improved system performance. Determining conditions under which greater visibility between collaborative agents will result in improved system performance is a topic for future research.

5 Conclusions

We have examined how changes in the accuracy of the leader's observation of the follower can affect the leader's value function. We have given conditions that insure improved observation quality improve the leader's value function, assuming the changes in observation quality do not cause the follower to change its policy. We demonstrated that when changes in observation quality cause the follower to change its policy, discontinuities in the leader's value function can result, as a function of observation quality, and that these discontinuities can be beneficial or not beneficial to the leader. We showed that when the two agents are collaborative, i.e., share the same reward structure, and the follower has complete visibility of the leader's initial conditions, discontinuities in the leader's value function do not occur. However, whether or not the agents in the POMG are collaborative, improved quality of the leader's observations of the follower do not necessarily lead to improved leader performance.

This research represents an initial investigation into the impact of observation quality on performance for the POMG under very specific assumptions (there are two agents, a leader and a follower, and the follower selects its policy with complete knowledge of the leader's policy selection) and with focus on how the leader's observation quality of the follower impacts the leader's value function. Future directions for research on the interplay between observation quality and control in the context of the POMG appear numerous.

Acknowledgments This material is based upon work supported by the US Department of Homeland Security under Grant Award Number 2010-ST-061-FD0001 through a grant awarded by the National Center for Food Protection and Defense at the University of Minnesota. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the US Department of Homeland Security or the National Center for Food Protection and Defense.

6 Appendix 1: Proofs

Proof of Theorem 1 Proof of Theorem 1 is given in White and Harrington (1980). □

Proof of Corollary 1 Proof of (a) follows from the fact that an optimal policy has a concave value function and can be found in White and Harrington (1980) and Zhang (2010). Proof of (b) follows from the facts that for any stochastic matrix Q', there are stochastic matrices R and R'' such that $Q'R = Q$ and $Q''R'' = Q'$. □

Proof of Proposition 1 Proof follows the proof of Proposition 2 in Chang et al. (2014). □

Proof of Proposition 2 Proof follows from Lemma 1 in Ortiz et al. (2013) and Proposition 2 in Chang et al. (2014). □

Proof of Corollary 2 (1) follows directly from Theorem 1 and Proposition 1. It is sufficient to show that $v^L(\pi^L, \pi^F, Q)(d^L(t,\tau), y^L(t))$ is concave in $y^L(t)$ for (2) to hold. It follows from Proposition 2 that

$$\begin{aligned} &v^L(\pi^L, \pi^F, Q)(d^L(t,\tau), y^L(t)) \\ &\quad = \max_{\rho^L} \sum_{d^F(t,\tau)} P(d^F(t,\tau)|d^L(t)) g^L(d(t,\tau), \rho^L, \pi^*(\rho^L, Q), Q), \end{aligned}$$

which is concave in $y^L(t)$. □

Proof of Proposition 3 For any scalar valued function v dependent on $(d(t,\tau))$, define

$$\begin{aligned} &[Hv](d(t,\tau)) = R^L(d(t,\tau), \pi^L, \pi^F) \\ &\quad + \beta \sum_{\varsigma(t+1)} P(\varsigma(t+1)|d(t,\tau), \pi^L, \pi^F, Q) \times v(\{\varsigma(t+1), d(t,\tau-1)\}). \end{aligned}$$

Define H' identically to H but replace Q by Q', where we note:

$$\begin{aligned} &P(\varsigma(t+1)|d(t,\tau), \pi^L, \pi^F, Q) \\ &\quad = \sum_{a(t)} P(a(t)|d(t,\tau)) P(z^L(t+1)|s^F(t+1)) P(z^F(t+1), s(t+1)|s(t), a(t)). \end{aligned}$$

Let g and g' be the fixed points of H and H', respectively [Existence and uniqueness of these fixed points are assured by Theorem 6.2.3 in Puterman (1994)]. Then,

$$\begin{aligned} &g(d(t,\tau)) - g'(d(t,\tau)) = [Hg](d(t,\tau)) - [H'g'](d(t,\tau)) - [Hg'](d(t,\tau)) + [Hg'](d(t,\tau)) \\ &\quad = \beta \sum_{a(t)} P(a(t)|d(t,\tau)) \times X, \end{aligned}$$

where

$$\begin{aligned} X = &\sum_{z(t+1)} \sum_{s(t+1)} P(z(t+1), s(t+1)|s(t), a(t)) \\ &\times \{g(\{\varsigma(t+1), d(t,\tau-1)\}) - g'(\{\varsigma(t+1), d(t,\tau-1)\})\} \\ &+ \sum_{z(t+1)} \sum_{s(t+1)} [P(z^L(t+1)|s^F(t+1)) - P'(z^L(t+1)|s^F(t+1))] P(z^F(t+1), \\ &\quad s(t+1)|s(t), a(t)) \times g'(\{\varsigma(t+1), d(t,\tau-1)\}) \end{aligned}$$

Note, $||g'|| \leq \frac{M}{1-\beta}$, where $M = \max_s \max_a r^L(s,a)$. Then, it is straightforward to show that

$$||g - g'|| \leq \beta ||g - g'|| + \beta ||Q - Q'|| \frac{M}{1-\beta}$$

and hence,

$$||g - g'|| \leq \frac{\beta M}{(1-\beta)^2} ||Q - Q'||.$$

□

Proof of Proposition 4 If Q is such that

$$v^F(\pi^L, \pi^F, Q)(d^F(0)) - v^F(\pi^L, \rho^F, Q)(d^F(0)) \geq 0$$

for all $\rho^F \neq \pi^F$, $\rho^F, \pi^F \in \Pi^F$ and given $d^F(0)$, then $Q \in K(\pi^L, \pi^F)$. Note

$$\begin{aligned} & v^F(\pi^L, \pi^F, Q)(d^F(0)) - v^F(\pi^L, \rho^F, Q)(d^F(0)) \\ &\quad = v^F(\pi^L, \pi^F, Q)(d^F(0)) - v^F(\pi^L, \pi^F, Q')(d^F(0)) + v^F(\pi^L, \rho^F, Q')(d^F(0)) \\ &\qquad - v^F(\pi^L, \rho^F, Q)(d^F(0)) + v^F(\pi^L, \pi^F, Q')(d^F(0)) - v^F(\pi^L, \rho^F, Q')(d^F(0)) \\ &\quad \geq -\frac{2\beta M}{(1-\beta)^2}||Q - Q'|| + b, \end{aligned}$$

where the inequality follows from Proposition 3 and definitions of b and M. The result follows from the fact that $b - \frac{2\beta M}{(1-\beta)^2}||Q - Q'|| \geq 0$ if and only if $||Q - Q'|| \leq \frac{b(1-\beta)^2}{2\beta M}$. □

Proof of Lemma 1 The proof follows from Proposition 4.7.3 (Puterman, p. 106) and a standard limit procedure. □

Proof of Proposition 5 Lemma 1 guarantees that $g^L(d^L(t,\tau), Q^*, \pi^L, \pi')$ for $\pi' \in \{\pi^F, \rho^F\}$ is isotone in $d(t,\tau)$. It follows from Lemma 4.7.2 (Puterman, p. 106) that

$$\begin{aligned} & \sum_{\varsigma(t+1)} P(\varsigma(t+1)|d(t,\tau), Q^*, \pi^L, \rho^F) \times g^L(d(t,\tau)|Q^*, \pi^L, \pi^F) \\ &\quad \geq \sum_{\varsigma(t+1)} P(\varsigma(t+1)|d(t,\tau), Q^*, \pi^L, \pi^F) \times g^L(d(t,\tau)|Q^*, \pi^L, \pi^F). \end{aligned}$$

Thus,,

$$\begin{aligned} & g^L(d(t,\tau), Q^*, \pi^L, \pi^F) \\ &\quad \leq R^L(d(t,\tau), \pi^L, \rho^F) + \beta \sum_{\varsigma(t+1)} P(\varsigma(t+1)|Q^*, \pi^L, \rho^F) \times g^L(d(t,\tau), Q^*, \pi^L, \pi^F). \end{aligned} \tag{6}$$

Let

$$\begin{aligned} & [Hv](d(t,\tau)) \\ &\quad = R^L(d(t,\tau), \pi^L, \rho^F) + \beta \sum_{\varsigma(t+1)} P(\varsigma(t+1)|d(t,\tau), Q^*, \pi^L, \rho^F) \\ &\quad \times v(\{\varsigma(t+1), d(t,\tau-1)\}). \end{aligned}$$

Define the sequence $\{v^n\}$ as $v^{n+1} = Hv^n$, where $v^0(d(t,\tau)) = g^L(d(t,\tau), Q^*, \pi^L, \pi^F)$. We remark that $\lim_{n\to\infty} ||v^n - v^*|| = 0$, where $v^*(d(t,\tau)) = g^L(d(t,\tau), Q^*, \pi^L, \rho^F)$. It is straightforward to show that $v \leq v'$ implies $Hv \leq Hv'$. Eq. (6) has shown $v^0 \leq v^1$. Lemma 1 guarantees that v^n is isotone in $d(t,\tau)$ for $n \geq 1$. Hence, by induction, $v^n \leq v^{n+1}$ and therefore $v^n \leq v^*$ for all n. Thus, $g^L(d(t,\tau), Q^*, \pi^L, \pi^F) \leq g^L(d(t,\tau), Q^*, \pi^L, \rho^F)$ for all $d(t,\tau)$ and hence $v^L(\pi^L, \pi^F, Q^*)(d^L(0)) \leq v^L(\pi^L, \rho^F, Q^*)(d^L(0))$. □

Proof of Proposition 6 Assumption (1) implies that $v^F(\pi^L, \pi', Q^*)(d^F(0)) = g^F(d(0,\tau), Q^*, \pi^L, \pi')$ for $\pi' \in \{\pi^F, \rho^F\}$. Assumption (2) implies $v^F(\pi^L, \pi^F, Q^*)(d^F(0)) = v^F(\pi^L, \rho^F, Q^*)(d^F(0))$. Hence, $g^F(d(0,\tau), \pi^L, \pi^F) = g^F(d(0,\tau), \pi^L, \rho^F)$. Assumption (3) implies $g^L = g^F$. □

7 Appendix 2: Parameters for the examples

The parameters in Example 1 are:

- transition probabilities:

$$P^L(1) = \begin{bmatrix} 0.6229 & 0.3771 \\ 0.7506 & 0.2494 \end{bmatrix}, \; P^L(2) = \begin{bmatrix} 0.7531 & 0.2469 \\ 0.1761 & 0.8239 \end{bmatrix}$$

$$P^F(1) = \begin{bmatrix} 0.2232 & 0.7768 \\ 0.5131 & 0.4869 \end{bmatrix}, \; P^F(2) = \begin{bmatrix} 0.9449 & 0.0551 \\ 0.2663 & 0.7337 \end{bmatrix}$$

- reward structure $r^k(s^L, s^F, a^L, a^F), k \in \{L, F\}$:

$$r^F = \begin{array}{c|cccc|} & (1^L,1^F)_a & (1^L,2^F)_a & (2^L,1^F)_a & (2^L,2^F)_a \\ (1^L,1^F)_s & 3.9855 & 1.2631 & 3.8957 & 4.9839 \\ (1^L,2^F)_s & 8.3138 & 8.6463 & 4.9014 & 5.2923 \\ (2^L,1^F)_s & 1.8500 & 7.6665 & 8.8690 & 9.0970 \\ (2^L,2^F)_s & 5.0079 & 5.6435 & 9.0505 & 5.7860 \end{array}$$

(1) example(a):

$$r^L = \begin{array}{c|cccc|} & (1^L,1^F)_a & (1^L,2^F)_a & (2^L,1^F)_a & (2^L,2^F)_a \\ (1^L,1^F)_s & 6.8156 & 8.2357 & 8.9439 & 9.5346 \\ (1^L,2^F)_s & 4.6326 & 1.7501 & 5.1656 & 5.4088 \\ (2^L,1^F)_s & 2.1216 & 1.6357 & 7.0270 & 6.7973 \\ (2^L,2^F)_s & 0.9852 & 6.6599 & 1.5359 & 0.3656 \end{array}$$

(2) example(b):

$$r^L = \begin{array}{c|cccc|} & (1^L,1^F)_a & (1^L,2^F)_a & (2^L,1^F)_a & (2^L,2^F)_a \\ (1^L,1^F)_s & 7.9466 & 7.5195 & 6.7120 & 3.9076 \\ (1^L,2^F)_s & 5.7739 & 2.2867 & 7.1521 & 8.1614 \\ (2^L,1^F)_s & 4.4004 & 0.6419 & 6.4206 & 3.1743 \\ (2^L,2^F)_s & 2.5761 & 7.6733 & 4.1905 & 8.1454 \end{array}$$

(3) example(c):

$$r^L = \begin{array}{c|cccc|} & (1^L,1^F)_a & (1^L,2^F)_a & (2^L,1^F)_a & (2^L,2^F)_a \\ (1^L,1^F)_s & 8.0549 & 8.8651 & 9.7868 & 0.5962 \\ (1^L,2^F)_s & 5.7672 & 0.2867 & 7.1269 & 6.8197 \\ (2^L,1^F)_s & 1.8292 & 4.8990 & 5.0047 & 0.4243 \\ (2^L,2^F)_s & 2.3993 & 1.6793 & 4.7109 & 0.7145 \end{array}$$

(4) example(d):

$$r^L = \begin{array}{c|cccc|} & (1^L,1^F)_a & (1^L,2^F)_a & (2^L,1^F)_a & (2^L,2^F)_a \\ (1^L,1^F)_s & 4.9417 & 8.9092 & 0.3054 & 9.0472 \\ (1^L,2^F)_s & 7.7905 & 3.3416 & 7.4407 & 6.0987 \\ (2^L,1^F)_s & 7.1504 & 6.9875 & 5.0002 & 6.1767 \\ (2^L,2^F)_s & 9.0372 & 1.9781 & 4.7992 & 8.5944 \end{array}$$

(5) example(e):

$$r^L = \begin{array}{c|cccc|} & (1^L,1^F)_a & (1^L,2^F)_a & (2^L,1^F)_a & (2^L,2^F)_a \\ (1^L,1^F)_s & 6.3114 & 9.9685 & 4.3000 & 0.6463 \\ (1^L,2^F)_s & 8.5932 & 5.5354 & 4.9181 & 4.3618 \\ (2^L,1^F)_s & 9.7422 & 5.1546 & 0.7104 & 8.2663 \\ (2^L,2^F)_s & 5.7084 & 3.3068 & 8.8774 & 3.9453 \end{array}$$

(6) example(f):

$$r^L = \begin{array}{c|cccc|} & (1^L,1^F)_a & (1^L,2^F)_a & (2^L,1^F)_a & (2^L,2^F)_a \\ (1^L,1^F)_s & 0.8348 & 8.9075 & 9.2831 & 8.6271 \\ (1^L,2^F)_s & 6.2596 & 9.8230 & 5.8009 & 4.8430 \\ (2^L,1^F)_s & 6.6094 & 7.6903 & 0.1698 & 8.4486 \\ (2^L,2^F)_s & 7.2975 & 5.8145 & 1.2086 & 2.0941 \end{array}$$

The parameters in Example 2 are:

- transition probabilities: $P(s(t+1), z^F(t+1)|s(t), \pi^L, \rho^F) =$

$$\begin{array}{c|cccccccc|} & 1^L_s,1^F_z,1^F_s & 1^L_s,1^F_z,2^F_s & 1^L_s,2^F_z,1^F_s & 1^L_s,2^F_z,2^F_s & 2^L_s,1^F_z,1^F_s & 2^L_s,1^F_z,2^F_s & 2^L_s,2^F_z,1^F_s & 2^L_s,2^F_z,2^F_s \\ (1^L,1^F)_s,(1^L,1^F)_a & 0.2371 & 0.1766 & 0.1741 & 0.1291 & 0.1268 & 0.0990 & 0.0479 & 0.0095 \\ (1^L,1^F)_s,(1^L,2^F)_a & 0.2371 & 0.1766 & 0.1741 & 0.1291 & 0.1268 & 0.0095 & 0.0479 & 0.0990 \\ (1^L,1^F)_s,(2^L,1^F)_a & 0.2371 & 0.1766 & 0.1741 & 0.1291 & 0.0990 & 0.0479 & 0.0095 & 0.1268 \\ (1^L,1^F)_s,(2^L,2^F)_a & 0.2371 & 0.1766 & 0.1741 & 0.1291 & 0.0479 & 0.0990 & 0.0095 & 0.1268 \\ (1^L,2^F)_s,(1^L,1^F)_a & 0.2371 & 0.1766 & 0.1741 & 0.1291 & 0.0095 & 0.0990 & 0.0479 & 0.1268 \\ (1^L,2^F)_s,(1^L,2^F)_a & 0.2371 & 0.1766 & 0.1741 & 0.1268 & 0.0095 & 0.0990 & 0.0479 & 0.1291 \\ (1^L,2^F)_s,(2^L,1^F)_a & 0.2371 & 0.1766 & 0.1741 & 0.0990 & 0.0095 & 0.1268 & 0.0479 & 0.1291 \\ (1^L,2^F)_s,(2^L,2^F)_a & 0.2371 & 0.1766 & 0.1268 & 0.1291 & 0.0095 & 0.0990 & 0.0479 & 0.1741 \\ (2^L,1^F)_s,(1^L,1^F)_a & 0.2371 & 0.1766 & 0.1268 & 0.0990 & 0.0095 & 0.1291 & 0.0479 & 0.1741 \\ (2^L,1^F)_s,(1^L,2^F)_a & 0.2371 & 0.1766 & 0.0990 & 0.1268 & 0.0095 & 0.1291 & 0.0479 & 0.1741 \\ (2^L,1^F)_s,(2^L,1^F)_a & 0.2371 & 0.1766 & 0.0479 & 0.1291 & 0.0095 & 0.1268 & 0.0990 & 0.1741 \\ (2^L,1^F)_s,(2^L,2^F)_a & 0.2371 & 0.1766 & 0.0479 & 0.1268 & 0.0095 & 0.1291 & 0.0990 & 0.1741 \\ (2^L,2^F)_s,(1^L,1^F)_a & 0.2371 & 0.1766 & 0.0095 & 0.1268 & 0.0479 & 0.1291 & 0.0990 & 0.1741 \\ (2^L,2^F)_s,(1^L,2^F)_a & 0.2371 & 0.1741 & 0.0095 & 0.1291 & 0.0479 & 0.1268 & 0.0990 & 0.1766 \\ (2^L,2^F)_s,(2^L,1^F)_a & 0.2371 & 0.1741 & 0.0095 & 0.1268 & 0.0479 & 0.1291 & 0.0990 & 0.1766 \\ (2^L,2^F)_s,(2^L,2^F)_a & 0.2371 & 0.1291 & 0.0095 & 0.1268 & 0.0479 & 0.1741 & 0.0990 & 0.1766 \end{array},$$

$P(s(t+1), z^F(t+1)|s(t), \pi^L, \pi^F) =$

$$\begin{array}{c|cccccccc|} & 1^L_s,1^F_z,1^F_s & 1^L_s,1^F_z,2^F_s & 1^L_s,2^F_z,1^F_s & 1^L_s,2^F_z,2^F_s & 2^L_s,1^F_z,1^F_s & 2^L_s,1^F_z,2^F_s & 2^L_s,2^F_z,1^F_s & 2^L_s,2^F_z,2^F_s \\ (1^L,1^F)_s,(1^L,1^F)_a & 0.21 & 0.17 & 0.18 & 0.12 & 0.13 & 0.07 & 0.05 & 0.07 \\ (1^L,1^F)_s,(1^L,2^F)_a & 0.21 & 0.17 & 0.18 & 0.12 & 0.13 & 0.02 & 0.04 & 0.13 \\ (1^L,1^F)_s,(2^L,1^F)_a & 0.21 & 0.17 & 0.18 & 0.12 & 0.11 & 0.05 & 0.01 & 0.15 \\ (1^L,1^F)_s,(2^L,2^F)_a & 0.21 & 0.17 & 0.18 & 0.12 & 0.06 & 0.10 & 0.01 & 0.15 \\ (1^L,2^F)_s,(1^L,1^F)_a & 0.21 & 0.17 & 0.18 & 0.12 & 0.02 & 0.10 & 0.05 & 0.15 \\ (1^L,2^F)_s,(1^L,2^F)_a & 0.21 & 0.17 & 0.18 & 0.11 & 0.02 & 0.10 & 0.04 & 0.17 \\ (1^L,2^F)_s,(2^L,1^F)_a & 0.21 & 0.17 & 0.18 & 0.10 & 0.01 & 0.11 & 0.04 & 0.18 \\ (1^L,2^F)_s,(2^L,2^F)_a & 0.21 & 0.17 & 0.14 & 0.12 & 0.01 & 0.10 & 0.05 & 0.20 \\ (2^L,1^F)_s,(1^L,1^F)_a & 0.21 & 0.17 & 0.14 & 0.09 & 0.01 & 0.13 & 0.05 & 0.20 \\ (2^L,1^F)_s,(1^L,2^F)_a & 0.21 & 0.17 & 0.11 & 0.12 & 0.01 & 0.13 & 0.05 & 0.20 \\ (2^L,1^F)_s,(2^L,1^F)_a & 0.21 & 0.17 & 0.06 & 0.13 & 0.01 & 0.12 & 0.10 & 0.20 \\ (2^L,1^F)_s,(2^L,2^F)_a & 0.21 & 0.17 & 0.06 & 0.12 & 0.01 & 0.13 & 0.10 & 0.20 \\ (2^L,2^F)_s,(1^L,1^F)_a & 0.21 & 0.17 & 0.02 & 0.12 & 0.05 & 0.13 & 0.10 & 0.20 \\ (2^L,2^F)_s,(1^L,2^F)_a & 0.21 & 0.16 & 0.02 & 0.12 & 0.045 & 0.13 & 0.10 & 0.215 \\ (2^L,2^F)_s,(2^L,1^F)_a & 0.21 & 0.16 & 0.02 & 0.11 & 0.04 & 0.135 & 0.10 & 0.225 \\ (2^L,2^F)_s,(2^L,2^F)_a & 0.21 & 0.13 & 0.01 & 0.13 & 0.03 & 0.165 & 0.10 & 0.225 \end{array}$$

$$Q^{L*} \in K(\pi^L, \pi^F) \cap K(\pi^L, \rho^F),\ Q^{L*} = \begin{bmatrix} 0.6 & 0.4 \\ 0.4 & 0.6 \end{bmatrix}$$

- reward structure:
$R^F(d^F(t,\tau), \pi^L, \rho^F) = [2.0944, -10, 9.1798, -10, 9.1768, -10, 9.1858, -10, -10, 9.3521, -10, 9.3522, -10, 9.3620, -10, 2.8030]$
$R^F(d^F(t,\tau), \pi^L, \pi^F) = [3.3540, 10, -9.3656, 10, -9.3656, 10, -9.3656, 10, 10, -9.3656, 10, -9.3656, 10, -9.3656, 10, -9.3656]$
$R^F(d^F(t,\tau), \rho^L, \rho^{F'}) = -\infty, \forall \rho^{F'} \in \Pi^F$
$R^L(d^F(t,\tau), \pi^L, \rho^F) = [2.9118, 2.8947, 2.8725, 2.8715, 2.7401, 2.7174, 2.4442, 2.4008, 1.8971, 1.6406, 1.4561, 0.8355, 0.4728, 0.4257, 0.3810, 0.2926]$
$R^L(d^F(t,\tau), \pi^L, \pi^F) = [2.0224, 1.9607, 1.9596, 1.9568, 1.9523, 1.9479, 1.7010, 1.6948, 1.2183, 0.9849, 0.8006, 0.4137, 0.3452, 0.2608, 0.2545, 0.0806]$

The parameters in Example 3 are:

(1) example(a):

- transition probabilities:

$$P^L(1) = \begin{bmatrix} 0.3202 & 0.6798 \\ 0.3044 & 0.6956 \end{bmatrix},\ P^L(2) = \begin{bmatrix} 0.7624 & 0.2376 \\ 0.3790 & 0.6210 \end{bmatrix}$$
$$P^F(1) = \begin{bmatrix} 0.2593 & 0.7407 \\ 0.6356 & 0.3644 \end{bmatrix},\ P^F(2) = \begin{bmatrix} 0.5221 & 0.4779 \\ 0.0994 & 0.9006 \end{bmatrix}$$

- reward structure $r^k(s^L, s^F, a^L, a^F), k \in \{L, F\}$:

$$r^F = r^L = \begin{array}{c} \\ (1^L,1^F)_s \\ (1^L,2^F)_s \\ (2^L,1^F)_s \\ (2^L,2^F)_s \end{array} \begin{array}{cccc} (1^L,1^F)_a & (1^L,2^F)_a & (2^L,1^F)_a & (2^L,2^F)_a \\ \left[\begin{matrix} 9.3942 \\ 4.1759 \\ 2.9319 \\ 0.0611 \end{matrix}\right. & \begin{matrix} 3.5084 \\ 5.0135 \\ 0.8495 \\ 1.7647 \end{matrix} & \begin{matrix} 6.4232 \\ 9.2925 \\ 8.3598 \\ 3.3969 \end{matrix} & \left.\begin{matrix} 0.2711 \\ 2.2993 \\ 6.4256 \\ 5.4417 \end{matrix}\right] \end{array}$$

(2) example(b):

- transition probabilities:

$$P^L(1) = \begin{bmatrix} 0.3277 & 0.6723 \\ 0.9623 & 0.0377 \end{bmatrix},\ P^L(2) = \begin{bmatrix} 0.4723 & 0.5277 \\ 0.4469 & 0.5531 \end{bmatrix}$$
$$P^F(1) = \begin{bmatrix} 0.8815 & 0.1185 \\ 0.6147 & 0.3853 \end{bmatrix},\ P^F(2) = \begin{bmatrix} 0.0641 & 0.9359 \\ 0.2062 & 0.7938 \end{bmatrix}$$

- reward structure $r^k(s^L, s^F, a^L, a^F), k \in \{L, F\}$:

$$r^F = r^L = \begin{array}{c} \\ (1^L,1^F)_s \\ (1^L,2^F)_s \\ (2^L,1^F)_s \\ (2^L,2^F)_s \end{array} \begin{array}{cccc} (1^L,1^F)_a & (1^L,2^F)_a & (2^L,1^F)_a & (2^L,2^F)_a \\ \left[\begin{matrix} 6.0012 \\ 9.8051 \\ 7.6433 \\ 7.1890 \end{matrix}\right. & \begin{matrix} 8.8066 \\ 0.1065 \\ 5.7154 \\ 9.7160 \end{matrix} & \begin{matrix} 5.3363 \\ 0.6272 \\ 5.6470 \\ 2.2813 \end{matrix} & \left.\begin{matrix} 4.4058 \\ 0.9891 \\ 1.2712 \\ 8.0163 \end{matrix}\right] \end{array}$$

(3) example(c):

- transition probabilities:

$$P^L(1) = \begin{bmatrix} 0.7657 & 0.2343 \\ 0.9270 & 0.0730 \end{bmatrix}, P^L(2) = \begin{bmatrix} 0.5570 & 0.4430 \\ 0.8113 & 0.1887 \end{bmatrix}$$

$$P^F(1) = \begin{bmatrix} 0.3594 & 0.6406 \\ 0.0007 & 0.9993 \end{bmatrix}, P^F(2) = \begin{bmatrix} 0.4647 & 0.5353 \\ 0.7964 & 0.2036 \end{bmatrix}$$

- reward structure $r^k(s^L, s^F, a^L, a^F), k \in \{L, F\}$:

$$r^F = r^L = \begin{array}{c|cccc|} & (1^L, 1^F)_a & (1^L, 2^F)_a & (2^L, 1^F)_a & (2^L, 2^F)_a \\ (1^L, 1^F)_s & 3.3740 & 8.2159 & 1.6372 & 3.5419 \\ (1^L, 2^F)_s & 7.5482 & 8.1603 & 5.8402 & 3.2979 \\ (2^L, 1^F)_s & 7.0649 & 8.1617 & 9.0912 & 7.2706 \\ (2^L, 2^F)_s & 4.4256 & 5.0106 & 2.8963 & 6.9836 \end{array}$$

(4) example(d):

- transition probabilities:

$$P^L(1) = \begin{bmatrix} 0.5983 & 0.4017 \\ 0.6592 & 0.3408 \end{bmatrix}, P^L(2) = \begin{bmatrix} 0.8785 & 0.1215 \\ 0.3068 & 0.6932 \end{bmatrix}$$

$$P^F(1) = \begin{bmatrix} 0.7036 & 0.2964 \\ 0.5885 & 0.4115 \end{bmatrix}, P^F(2) = \begin{bmatrix} 0.0593 & 0.9407 \\ 0.3593 & 0.6407 \end{bmatrix}$$

- reward structure $r^k(s^L, s^F, a^L, a^F), k \in \{L, F\}$:

$$r^F = r^L = \begin{array}{c|cccc|} & (1^L, 1^F)_a & (1^L, 2^F)_a & (2^L, 1^F)_a & (2^L, 2^F)_a \\ (1^L, 1^F)_s & 7.3817 & 6.7738 & 4.4108 & 0.6868 \\ (1^L, 2^F)_s & 9.0358 & 7.3192 & 3.5113 & 3.4176 \\ (2^L, 1^F)_s & 7.2087 & 8.9864 & 9.3087 & 0.1376 \\ (2^L, 2^F)_s & 6.4387 & 9.3377 & 9.5397 & 3.9611 \end{array}$$

References

Bassan, B., Gossner, O., Scarsini, M., & Zamir, S. (2003). Positive value of information in games. *Internal Journal of Game Theory*, *32*, 17–31.

Bertsekas, D. P. (1976). *Dynamic programming and stochastic control*. New York: Academic Press.

Bernstein, D. S., Givan, R., Immerman, N., & Zilberstein, S. (2002). The complexity of decentralized control of Markov decision processes. *Mathematics of Operations Research*, *27*(4), 819–840.

Bier, V. M., Oliveros, S., & Samuelson, L. (2007). Choosing what to protect: Strategic defensive allocation against an unknown attacker. *Journal of Public Economic Theory*, *9*(4), 563–587.

Cassandra, A. R., Kaelbling, L. P., & Littman, M. L. (1994). Acting optimally in partially observable stochastic domains. In *Proceedings twelfth national conference on artificial intelligence (AAAI-94)* (pp. 1023–1028). WA: Seattle.

Cassandra, A. R., Littman, M. L., & Zhang, N. L. (1997). Incremental pruning: A simple, fast, exact method for partially observable Markov decision processes. In *Proceedings thirteenth annual conference on uncertainty in artificial intelligence (UAI-97)* (pp. 54–61). San Francisco, CA: Morgan Kaufmann.

Chang, Y. L., Erera, A.L., & White, C. C. (2014). A leader–follower partially observed multiobjective Markov game, submitted for publication.

Chu, W. H. J., & Lee, C. C. (2006). Strategic information sharing in a supply chain. *European Journal of Operational Reserch*, *174*, 1567–1579.

Ezell, B. C., Bennett, S. P., von Winterfeldt, D., Sokolowski, J., & Collins, A. J. (2010). Probabilistic risk analysis and terrorism risk. *Risk Analysis*, *30*(4), 575–589.

Hansen, E. A., Bernstein, D. S., & Zilberstein, S. (2004). Dynamic programming for partially observable stochastic games. In *Proceedings of the nineteenth national conference on artificial intelligence* (pp. 709–715). San Jose: California.

Kamien, M. I., Tauman, Y., & Zamir, S. (1990). On the value of information in a strategic conflict. *Games and Economic Behavior*, *2*, 129–153.

Kumar, A., & Zilberstein, S. (2009). Dynamic programming approximations for partially observable stochastic games. In *Proceedings of the twenty-second international FLAIRS conference* (pp. 547–552). Florida: Sanibel Island.

Lehrer, E., & Rosenberg, D. (2006). What restrictions do Bayesian games impose on the value of information? *Journal of Mathematical Economics*, *42*, 343–357.

Lehrer, E., & Rosenberg, D. (2010). A note on the evaluation of information in zero-sum repeated games. *Journal of Mathematical Economics*, *46*, 393–399.

Leng, M. M., & Parlar, M. (2009). Allocation of cost savings in a three-level supply chain with demand information sharing: A cooperate-game approach. *Operations Research*, *57*(1), 200–213.

Li, L. (2002). Information sharing in a supply chain with horizontal competition. *Management Science*, *48*(9), 1196–1212.

Lin, A. Z.-Z., Bean, J., & White, C. C. (1998). Genetic algorithm heuristics for finite horizon partially observed Markov decision problems, *Technical report*, University of Michigan, Ann Arbor.

Lin, A. Z.-Z., Bean, J., & White, C. C. (2004). A hybrid genetic/optimization algorithm for finite horizon partially observed Markov decision processes. *Journal on Computing*, *16*(1), 27–38.

Littman, M. L. (1994). The Witness algorithm: solving partially observable Markov decision processes, Brown University, Department of Computer Science, *Technical report*, CS-94-40.

Lovejoy, W. S. (1991). A survey of algorithmic methods for partially observed Markov decision process. *Annals of Operations Research*, *28*(1), 47–65.

Meuleau, N., Peshkin, L., Kim, K., & Kaelbling, L. P. (1999). Learning finite-state controllers for partially observable environments. In *Proceedings of the fifteenth conference on uncertainty in artificial intelligence* (pp. 427–436). Morgan Kaufmann Publishers.

Meyer, B. D., Lehrer, E., & Rosenberg, D. (2010). Evaluating information in zero-sum games with incomplete information on both sides. *Mathematics of Operations Research*, *35*(4), 851–863.

Monahan, G. E. (1982). A survey of partially observable Markov decision processes: Theory, models, and algorithms. *Management Science*, *28*, 1–16.

Ortiz, O. L., Erera, A. L., & White, C. C. (2013). State observation accuracy and finite-memory policy performance. *Operations Research Letters*, *41*, 477–481.

Platzman, L. K. (1977). *Finite memory estimation and control of finite probabilistic systems*, PhD thesis, Cambridge, MA: Massachusetts Institute of Technology.

Platzman, L. K. (1980). Optimal infinite-horizon undiscounted control of finite probabilistic systems. *SIAM Journal on Control and Optimization*, *18*, 362–380.

Poupart, P., & Boutilier, C. (2004). *Bounded finite state controllers, Advances in Neural Information Processing Systems, 16*. Cambridge, MA: MIT Press.

Puterman, M. L. (1994). *Markov decision processes: Discrete dynamic programming*. New York: Wiley.

Rabinovich, Z., Goldman, C. V., & Rosenschein, J. S. (2003). The complexity of multiagent systems: The price of silence. *Proceedings of the second international joint conference on autonomous agents and multi-agent systems (AAMAS)* (pp. 1102–1103). Australia, Melbourne.

Shapley, L. S. (1953). Stochastic games, *Proceedings of the national academy of sciences of the USA*, *39*, 1095–1100.

Smallwood, R. D., & Sondik, E. J. (1973). The optimal control of partially observable Markov decision processes over a finite horizon. *Operations Research*, *21*, 1071–1088.

Sondik, E. J. (1978). The optimal control of partially observable Markov processes over the infinite horizon: Discounted costs. *Operations Research*, *26*, 282–304.

Wakker, P. (1988). Nonexpected utility as aversion of information. *Journal of Behavioral Decision Making*, *1*, 169–175.

White, C. C., & Harrington, D. P. (1980). Application of Jensen's inequality to adaptive suboptimal design. *Journal of Optimization Theory and Application*, *32*, 89–99.

White, C. C., & Scherer, W. T. (1989). Solution procedures for partially observed Markov decision processes. *Operations Research*, *37*, 791–797.

White, C. C. (1991). A survey of solution techniques for the partially observed Markov decision process. *Annals of Operations Research*, *32*, 215–230.

White, C. C., & Scherer, W. T. (1994). Finite-memory suboptimal design for partially observed Markov decision processes. *Operations Research*, *42*, 439–455.

Zhang, H. (2010). Partially observable Markov decision processes: A geometric technique and analysis. *Operations Research*, *58*, 214–228.

Zhuang, J., & Bier, V. M. (2010). Reasons for secrecy and deception in homeland-security resource allocation. *Risk Analysis*, *30*(12), 1737–1743.

Zhuang, J., Bier, V. M., & Alagoz, O. (2010). Modeling secrecy and deception in a multiple-period attacker-defender signaling game. *European Journal of Operational Research*, *203*, 409–418.