UIC AI Lab Technical Report

UIC-AIL-TR05

A Framework for Sequential Planning in Multi-Agent Settings

Piotr Gmytrasiewicz and **Prashant Doshi**
Artificial Intelligence Laboratory
Department of Computer Science
University of Illinois at Chicago, IL 60607

A Framework for Sequential Planning in Multi-Agent Settings

Piotr J. Gmytrasiewicz and Prashant Doshi*
Department of Computer Science
University of Illinois at Chicago
{piotr,pdoshi}@cs.uic.edu

July 20, 2004

Abstract

This paper extends the framework of partially observable Markov decision processes (POMPDs) to multi-agent settings by incorporating the notion of agent models into the state space. Agents maintain beliefs over physical states of the environment and over models of other agents and they use Bayesian update to maintain their beliefs over time. The solutions map belief states to actions. The models of other agents may include their belief states, and are related to agent types defined in games of incomplete information. We express the agents' autonomy by postulating that their models are not directly manipulable or observable by other agents. We show that a number of important properties of POMDPs carry over to our framework. Our approach complements a more traditional approach in stochastic games which uses equilibria as a solution paradigm. Our work seeks to avoid some of the drawbacks of game-theoretic equilibria which may be non-unique and are not able to capture off-equilibrium behaviors. We do so at the cost of having to represent, process and continually revise models of other agents. Agents' beliefs may be arbitrarily nested and optimal solutions to decision making problems are only asymptotically computable. However, approximate belief updates and approximately optimal plans are computable.

1 Introduction

We develop a framework for sequential rationality of autonomous agents interacting with other agents within a common, and possibly uncertain, environment. We use the normative paradigm of decision-theoretic planning under uncertainty as represented by partially observable Markov decision processes (POMDPs) [6, 18, 28]. We make this framework applicable to agents interacting with other agents by allowing them to have beliefs not only about the physical environment, but also about the other agents; i.e., their abilities, sensing capabilities, beliefs, preferences, and intentions.

The formalism of Markov decision processes has been extended to multiple agents before, giving rise to stochastic games or Markov games [12]. Traditionally, the solution concept used for stochastic games is that of Nash equilibria. Some recent work in AI follows that tradition [5, 16, 20, 21]. However, while Nash equilibria are useful for describing a multi-agent system when, and if, it has reached a stable state, this solution concept is not sufficient as a general control paradigm. The main reasons are that there may be multiple equilibria with no clear way to choose among them (non-uniqueness), and the fact that equilibria do not specify actions in cases in which agents believe that other agents may act not according to their equilibrium strategies (incompleteness) [4, 17].

Other extensions of POMDPs to multiple agents appeared in [3, 27]. They have been called decentralized POMDPs (DEC-POMDPs), and are related to decentralized control problems [26]. DEC-POMDP framework assumes that the agents are fully cooperative, i.e., they have common reward function and form a team. Furthermore, it is assumed that the optimal joint solution is computed centrally and then distributed among the agents, which makes it a variant of multi-body planning [28].

The main idea behind our formalism, called interactive POMDPs (I-POMDPs), is to allow an agent to have beliefs about the physical environment, as in POMDPs, but also over the other agent(s) in terms of their

*This research is supported by the National Science Foundation CAREER award IRI-9702132, and NSF award IRI-0119270.

preferences, capabilities, and beliefs. As such, the beliefs about others' beliefs, and their beliefs about others can be nested to infinite levels. I-POMDPs are applicable to autonomous self-interested agents with possibly conflicting objectives, operating in partially observable environments, who locally compute what actions they should execute to optimize their preferences given what they believe. We are motivated by a subjective approach to probability in games [17], and we combine POMDPs, Bayesian games, work on interactive belief systems [1, 10, 9, 14, 22], and insights from research on learning in game theory [11]. This approach, also called decision-theoretic approach to game theory [24], consists of predicting actions of other agents given all available information, and then of choosing the agent's own action [17]. This complements the traditional equilibrium approach [4, 17] in that it is centered on optimality and best response to anticipated action of other agent(s), rather then on stability. The question of whether, under what circumstances, and what kind of equilibria could arise from solutions to I-POMDPs is open.

Our approach avoids the difficulties of non-uniqueness and incompleteness of traditional equilibrium approach, but at the cost of processing and maintaining possibly infinitely nested interactive beliefs. Consequently, only approximate belief updates and approximately optimal solutions to planning problems are computable in general. We define a class of finitely nested I-POMDPs which are natural approximations of infinitely nested ones. We show that a number of properties that facilitate solutions of POMDPs carry over to I-POMDPs. In particular, the interactive beliefs are sufficient statistics for the histories of agent's observations, the belief update is a generalization of the update in POMDPs, the value function is piece-wise linear and convex in finitely nested beliefs, and the value iteration algorithm converges.

2 Background: Partially Observable Markov Decision Processes

A partially observable Markov decision processes (POMDP) [6, 15, 18, 23] of an agent i is defined as

$$\mathit{POMDP}_i = \langle S, A_i, T_i, \Omega_i, O_i, R_i \rangle \tag{1}$$

where: S is a set of possible states of the environment, A_i is a set of actions agent i can execute, T_i is a transition function – $T_i : S \times A_i \times S \to [0,1]$ which describes results of agent i's actions, Ω_i is the set of observations that the agent i can make, O_i is the agent's observation function – $O_i : S \times A_i \times \Omega_i \to [0,1]$ which specifies probabilities of observations if agent executes various actions that result in different states, R_i is the reward function representing the agent i's preferences – $R_i : S \times A_i \to \mathbf{R}$.

The belief update and the optimal solutions (depending on an optimality criterion) to POMDPs are described in the literature [6, 15, 18, 23].

2.1 Optimality Criteria

The agent's optimality criterion, OC_i, is needed to specify how rewards acquired over time are handled. Commonly used criteria include:

- A finite horizon criterion, in which the agent maximizes the expected value of the sum of first T rewards: $E(\sum_{t=0}^{T} r_t)$, where r_t is a reward obtained at time t and T is the length of the horizon. We will denote this criterion as fh^T.
- An infinite horizon criterion with discounting, according to which the agent maximizes $E(\sum_{t=0}^{\infty} \gamma^t r_t)$, where $0 < \gamma < 1$ is a discount factor. We will denote this criterion as ih^γ.
- An infinite horizon criterion with averaging, according to which the agent maximizes the average reward per time step. We will denote this as ih^{AV}.

In what follows, we concentrate on infinite horizon criterion with discounting, but our approach can be easily adapted to other criteria.

3 Agent Types and Frames

The POMDP definition above includes parameters that permit us to compute an agent's optimal behavior, conditioned on its beliefs. Let us collect these implementation independent factors into a construct we call an agent i's *type*.

Definition 1. (Type) *A type of an agent i is $\theta_i = \langle b_i, A_i, \Omega_i, T_i, O_i, R_i, OC_i \rangle$, where b_i is agent i's state of belief (an element of $\Delta(S)$), OC_i is its optimality criterion, and the rest of the elements are as defined before. Let Θ_i be the set of agent i's types.*

Given type, θ_i, and under the assumption that the agent is Bayesian-rational, the set of agent's optimal actions will be denoted as $OPT(\theta_i)$. In the next section, we generalize the notion of type to situations which include interactions with other agents; it then coincides with the notion of type used in Bayesian games [12, 14].

It is convenient to define the notion of a *frame*, $\widehat{\theta}_i$, of agent i:

Definition 2. (Frame) *A frame of an agent i is $\widehat{\theta}_i = \langle A_i, \Omega_i, T_i, O_i, R_i, OC_i \rangle$. Let $\widehat{\Theta}_i$ be the set of agent i's frames.*

For brevity one can write a type as consisting of an agent's belief together with its frame: $\theta_i = \langle b_i, \widehat{\theta}_i \rangle$.

4 Interactive POMDPs

As we mentioned, our intention is to generalize POMDPs to handle presence of other agents. We do this by including descriptions of other agents (their types for example) in the state space. For simplicity of presentation, we consider an agent i, that is interacting with one other agent, j. The formalization easily generalizes for larger number of agents.

Definition 3. (I-POMDP) *An* interactive POMDP *of agent i, I-POMDP$_i$, is:*

$$I\text{-}POMDP_i = \langle IS_i, A, T_i, \Omega_i, O_i, R_i \rangle \tag{2}$$

where:

• IS_i is a set of **interactive** states defined as $IS_i = S \times M_j$, where S is the set of states of the physical environment, and M_j is the set of possible models of agent j. Each model, $m_j \in M_j$, is a pair $m_j = \langle h_j, f_j \rangle$, where $f_j : H_j \rightarrow \Delta(A_j)$ is agent j's function, assumed computable, which maps possible histories of j's observations to distributions over its actions. h_j is an element of H_j.

Simple examples of models include a no-information model and ficticious play model, both of which are history independent. A no-information model [13] assume that each of the other agent's actions is executed with equal probability. Ficticious play [11] assumes that the other agent chooses according to a sequence of i.i.d. multinomial random variables with a fixed but unknown distribution, and the original agent's prior belief over that distribution takes a form of a Dirichlet distribution.

An important class of models, called intentional models, ascribe to the other agent beliefs and preferences. Intentional models are thus j's types, $\theta_j = \langle b_j, \widehat{\theta}_j \rangle$, together with the assumption that agent j is Bayesian rational. Agent j's belief is a probability distribution over the states of the world and the models of the agent i; $b_j \in \Delta(S \times M_i)$. The notion of a type we use here coincides with the notion of type in game theory, where it is defined as consisting of all of the agent i's private information relevant to its decision making [12, 14]. In particular, if agents' beliefs are private information, then their types involve possibly infinitely nested beliefs over others' types and their beliefs about others [1, 2, 8, 22].[1] They are related to recursive model structures in our prior work [13]. The definition of interactive state space is consistent with the notion of a completely specified state space put forward by Aumann [1]. Similar state spaces have been proposed in [8, 22].

• $A = A_i \times A_j$ is the set of joint moves of all agents.

[1] Implicit in the definition of nested beliefs is the assumption of coherency [8].

• T_i is the transition model. The usual way to define the transition probabilities in POMDPs is to assume that the agent's actions can change any aspect of the state description. In case of I-POMDPs, this would mean actions modifying any aspect of the interactive states, including other agents' observation histories or their functions, or, if they are modeled intentionally, their beliefs or reward functions. Allowing agents to directly manipulate other agents in such ways, however, violates the notion of agents' autonomy. Thus, we make the following simplifying assumption:

Model Non-manipulability Assumption (MNM): Agents' actions do not change the other agents' models directly.

Given this simplification, the transition model can be defined as $T_i : S \times A \times S \rightarrow [0, 1]$

Autonomy and the MNM assumption preclude direct "mind control" and implies that other agents' belief states can be changed only indirectly, typically by changing the environment in a way observable to them. In other words, agents' beliefs do change, like in POMDPs, but as a result of belief update after an observation, not as a direct result of any of the agents' actions.[2]

• Ω_i is defined as before in POMDP model.

• O_i is an observation function. In defining this function we make the following assumption:

Model Non-observability (MNO): Agents cannot observe other's models directly.

Given this assumption the observation function is defined as $O_i : S \times A \times \Omega_i \rightarrow [0, 1]$.

The MNO assumption expresses another aspect of autonomy – agents are autonomous in that their observations or functions, or beliefs and preferences in intentional models, are private and the other agents cannot observe them directly.[3]

• R_i is defined as $R_i : IS_i \times A \rightarrow \mathbf{R}$. We allow the agent to have preferences over physical states and models of other agents, but usually only the physical state will matter.

4.1 Belief Update in *I-POMDPs*

We will show that, as in POMDPs, agent's beliefs over their interactive states are *sufficient statistics*, i.e., they fully summarize the agent's observable histories. Further, we need to show how beliefs are updated after agent's action and observation, and how solution is defined. There are two differences that complicate belief update when compared to POMDPs. First, since the state of the physical environment depends on the actions performed by both agents the prediction of how the physical state changes has to be made based on the predicted actions of the other agent. The probabilities of other's actions are obtained based on their models. Thus, unlike in Bayesian and stochastic games, we do not assume that actions are fully observable by other agents. Rather, agents can attempt to infer what actions other agents have performed by sensing their results on the environment.Second, changes in the models of the other agents have to be included in the update. These reflect the update of the other agent's beliefs due to its new observation. In other words, the agent has to update its beliefs based on what it anticipates the other agent observes and how it updates. As could be expected, the update of the possibly infinitely nested belief over other's types is, in general, only asymptotically computable.

Proposition 1. (Sufficiency) *In an interactive POMDP of agent i, i's current belief, i.e., the probability distribution over the set $S \times \Theta_j$, is a sufficient statistic for the past history of i's observations.*

The next proposition defines the agent i's belief update function, $b_i^t(is^t) = Pr(is^t|o_i^t, a_i^{t-1}, b_i^{t-1})$, where $is^t \in IS_i$ is an interactive state. We will use the belief state estimation function, SE_{θ_i}, as an abbreviation for belief updates for individual states so that $b_i^t = SE_{\theta_i}(b_i^{t-1}, a_i^{t-1}, o_i^t)$. $\tau_{\theta_i}(b_i^{t-1}, a_i^{t-1}, o_i^t, b_i^t)$ will stand for $Pr(b_i^t|b_i^{t-1}, a_i^{t-1}, o_i^t)$. We will also define the set of type-dependent optimal actions of an agent, $OPT(\theta_i)$.

Proposition 2. (Belief Update) *Under the MNM and MNO assumptions, the belief update function for*

[2]The possibility that agents can influence the observational capabilities of other agents can be accommodated by including the factors that can change sensing capabilities in the set S.

[3]The possibility that agents can observe factors that may influence the observational capabilities of other agents is allowed by including these factors in S.

interactive POMDP $\langle IS_i, A, T_i, \Omega_i, O_i, R_i\rangle$ *is:*

$$\begin{array}{ll} b_i^t(is^t) & = \beta \sum_{is^{t-1}:\widehat{\theta}_j^{t-1}=\widehat{\theta}_j^t} b_i^{t-1}(is^{t-1}) \sum_{a_j^{t-1}} Pr(a_j^{t-1}|\theta_j^{t-1}) O_i(s^t, a^{t-1}, o_i^t) \sum_{o_j^t} \tau_{\theta_j^t}(b_j^{t-1}, a_j^{t-1}, o_j^t, b_j^t) \\ & \times O_j(s^t, a^{t-1}, o_j^t) T_i(s^{t-1}, a^{t-1}, s^t) \end{array} \tag{3}$$

where $is = (s, \theta_j)$, $is_j = (s, \theta_i)$, b_j^{t-1} and b_j^t are the belief elements of θ_j^{t-1} and θ_j^t, respectively, β is a normalizing constant, O_j is the observation function in θ_j^t, and $Pr(a_j^{t-1}|\theta_j^{t-1})$ is the probability that a_j^{t-1} is Bayesian rational for agent described by type θ_j^{t-1}. This probability is equal to $\frac{1}{|OPT(\theta_j)|}$ if $a_j^{t-1} \in OPT(\theta_j)$, and it is equal to zero otherwise. We define OPT in Section 4.2. [4]

Proof of Propositions 1 and 2. We start with Proposition 2, by applying the Bayes Theorem:

$$\begin{array}{l} b_i^t(is^t) = Pr(is^t|o_i^t, a_i^{t-1}, b_i^{t-1}) = \frac{Pr(is^t, o_i^t|a_i^{t-1}, b_i^{t-1})}{Pr(o_i^t|a_i^{t-1}, b_i^{t-1})} \\ \\ = \beta \sum_{is^{t-1}} b_i^{t-1}(is^{t-1}) Pr(is^t, o_i^t|a_i^{t-1}, is^{t-1}) \\ = \beta \sum_{is^{t-1}} b_i^{t-1}(is^{t-1}) \sum_{a_j^{t-1}} Pr(is^t, o_i^t|a_i^{t-1}, a_j^{t-1}, is^{t-1}) Pr(a_j^{t-1}|a_i^{t-1}, is^{t-1}) \\ = \beta \sum_{is^{t-1}} b_i^{t-1}(is^{t-1}) \sum_{a_j^{t-1}} Pr(is^t, o_i^t|a_i^{t-1}, a_j^{t-1}, is^{t-1}) Pr(a_j^{t-1}|is^{t-1}) \\ = \beta \sum_{is^{t-1}} b_i^{t-1}(is^{t-1}) \sum_{a_j^{t-1}} Pr(a_j^{t-1}|\theta_j^{t-1}) Pr(o_t^i|is^t, a^{t-1}, is^{t-1}) Pr(is^t|a^{t-1}, is^{t-1}) \\ = \beta \sum_{is^{t-1}} b_i^{t-1}(is^{t-1}) \sum_{a_j^{t-1}} Pr(a_j^{t-1}|\theta_j^{t-1}) Pr(o_t^i|is^t, a^{t-1}) Pr(is^t|a^{t-1}, is^{t-1}) \\ = \beta \sum_{is^{t-1}} b_i^{t-1}(is^{t-1}) \sum_{a_j^{t-1}} Pr(a_j^{t-1}|\theta_j^{t-1}) O_i(s^t, a^{t-1}, o_i^t) Pr(is^t|a^{t-1}, is^{t-1}) \end{array} \tag{4}$$

To simplify the term $Pr(is^t|a^{t-1}, is^{t-1})$ let us substitute the interactive state is^t with its components, $is^t = (s^t, \theta_j^t) = (s^t, b_j^t, \widehat{\theta}_j^t)$.

$$\begin{array}{ll} Pr(is^t|a^{t-1}, is^{t-1}) & = Pr(s^t, b_j^t, \widehat{\theta}_j^t|a^{t-1}, is^{t-1}) \\ & = Pr(b_j^t|s^t, \widehat{\theta}_j^t, a^{t-1}, is^{t-1}) Pr(s^t, \widehat{\theta}_j^t|a^{t-1}, is^{t-1}) \\ & = Pr(b_j^t|s^t, \widehat{\theta}_j^t, a^{t-1}, is^{t-1}) Pr(\widehat{\theta}_j^t|s^t, a^{t-1}, is^{t-1}) Pr(s^t|a^{t-1}, is^{t-1}) \\ & = Pr(b_j^t|s^t, \widehat{\theta}_j^t, a^{t-1}, is^{t-1}) I(\widehat{\theta}_j^{t-1}, \widehat{\theta}_j^t) T_i(s^{t-1}, a^{t-1}, s^t) \end{array} \tag{5}$$

The joint action pair, a^{t-1}, may change the physical state. The third term on the right-hand side above captures this transition. We utilized the MNM assumption to replace the second term with a boolean identity function, $I(\widehat{\theta}_j^{t-1}, \widehat{\theta}_j^t)$, which equals 1 if the two frames are identical, and 0 otherwise. The third term can be expressed using the transition function. Let us turn our attention to the first term,

$$\begin{array}{ll} Pr(b_j^t|s^t, \widehat{\theta}_j^t, a^{t-1}, is^{t-1}) & = \sum_{o_j^t} Pr(b_j^t|s^t, \widehat{\theta}_j^t, a^{t-1}, is^{t-1}, o_j^t) Pr(o_j^t|s^t, \widehat{\theta}_j^t, a^{t-1}, is^{t-1}) \\ & = \sum_{o_j^t} Pr(b_j^t|s^t, \widehat{\theta}_j^t, a^{t-1}, is^{t-1}, o_j^t) Pr(o_j^t|s^t, \widehat{\theta}_j^t, a^{t-1}) \\ & = \sum_{o_j^t} \tau_{\theta_j^t}(b_j^{t-1}, a_j^{t-1}, o_j^t, b_j^t) O_j(s_t, a^{t-1}, o_j^t) \end{array} \tag{6}$$

In the above equation, the first term on the right-hand side is 1 if agent j's belief update, $SE_{\theta_j}(b_j^{t-1}, a_j^{t-1}, o_j^t)$ generates a belief state equal to b_j^t. In the second term of the right-hand side, the MNO assumption makes it possible to replace $Pr(o_j^t|s^t, \widehat{\theta}_j^t, a^{t-1})$ with $O_j(s^t, a^{t-1}, o_j^t)$.

Let us now substitute Eq. 6 into Eq. 5.

$$Pr(is^t|a^{t-1}, is^{t-1}) = \sum_{o_j^t} \tau_{\theta_j^t}(b_j^{t-1}, a_j^{t-1}, o_j^t, b_j^t) O_j(s^t, a^{t-1}, o_j^t) I(\widehat{\theta}_j^{t-1}, \widehat{\theta}_j^t) T_i(s^{t-1}, a^{t-1}, s^t) \tag{7}$$

The final equation for belief update results by replacing Eq. 7 into Eq. 4.

[4] If the agent's prior belief over IS_i is given by a probability density function then the $\sum_{is^{t-1}}$ is replaced by an integral.

$$\begin{aligned} b_i^t(is^t) \;&= \beta \sum_{is^{t-1}} b_i^{t-1}(is^{t-1}) \sum_{a_j^{t-1}} Pr(a_j^{t-1}|\theta_j^{t-1}) O_i(s^t, a^{t-1}, o_i^t) \sum_{o_j^t} \tau_{\theta_j^t}(b_j^{t-1}, a_j^{t-1}, o_j^t, b_j^t) \\ &\times O_j(s^t, a^{t-1}, o_j^t) I(\widehat{\theta}_j^{t-1}, \widehat{\theta}_j^t) T_i(s^{t-1}, a^{t-1}, s^t) \end{aligned}$$

We can simply the above equation by removing the term $I(\widehat{\theta}_j^{t-1}, \widehat{\theta}_j^t)$ in the following manner:

$$\begin{aligned} b_i^t(is^t) \;&= \beta \sum_{is^{t-1}|\widehat{\theta}_j^{t-1}=\widehat{\theta}_j^t} b_i^{t-1}(is^{t-1}) \sum_{a_j^{t-1}} Pr(a_j^{t-1}|\theta_j^{t-1}) O_i(s^t, a^{t-1}, o_i^t) \sum_{o_j^t} \tau_{\theta_j^t}(b_j^{t-1}, a_j^{t-1}, o_j^t, b_j^t) \\ &\times O_j(s^t, a^{t-1}, o_j^t) T_i(s^{t-1}, a^{t-1}, s^t) \end{aligned}$$

Since proposition 2 expresses the belief $b_i^t(is^t)$ in terms of parameters of the previous time step only, Proposition 1 holds as well. □

Intuitively, proposition 1 holds since, as in POMDPs [29], all of the dynamic elements (state transitions and observations) of the model depend on the previous state, not on prior observations and actions.

Proposition 2 and Eq. 3 have a lot in common with belief update in POMDPs, as should be expected. Both depend on agent i's observation and transition functions. However, since agent i's observations also depend on agent j's actions, the probabilities of various actions of j have to be included (in the first line of Eq. 3.) Further, since the update of agent j's model depends on what j observes, the probabilities of various observations have to be included (in the second line of Eq. 3.) The update of j's beliefs is represented by the term $\tau_{\hat{\theta}_j}$ term.

Since the agent i's beliefs could be infinitely nested, the belief update in I-POMDPs can, in general, be calculated only asymptotically. The belief update can easily be generalized to the setting where more than one other agents co-exist with agent i.

4.2 Value Function and Solutions in I-POMDPs

Analogously to POMDPs, each belief state in I-POMDP has an associated value reflecting the maximum payoff the agent can expect in this belief state:

$$U(\theta_i) = \max_{a_i \in A_i} \left\{ \sum_{is} ER_i^{a_j}(is, a_i) b_i(is) + \gamma \sum_{o_i \in \Omega_i} Pr(o_i|a_i, b_i) U(\langle SE_{\theta_i}(b_i, a_i, o_i), \widehat{\theta}_i \rangle) \right\} \tag{8}$$

where, $ER_i^{a_j}(is, a_i) = \sum_{a_j} R_i(is, a_i, a_j) Pr(a_j|\theta_j)$ (since $is = (s, \theta_j)$). Eq. 8 defines value iteration in I-POMDPs.

Agent i's optimal action, a^*, for the case of infinite horizon criterion with discounting, is an element of the set of optimal actions for the belief state, $OPT(\theta_i)$, defined as:

$$OPT(\theta_i) = \underset{a_i \in A_i}{argmax} \left\{ \sum_{is} ER_i^{a_j}(is, a_i) b_i(is) + \gamma \sum_{o_i \in \Omega_i} Pr(o_i|a_i, b_i) U(\langle SE_{\theta_i}(b_i, a_i, o_i), \widehat{\theta}_i \rangle) \right\} \tag{9}$$

As in the case of belief update, due to possibly infinitely nested beliefs, a step of value iteration and optimal actions are only asymptotically computable.

5 Finitely Nested I-POMDPs

Possibly infinite nesting of agents' beliefs in intentional models presents an obvious obstacle to computing the belief updates and optimal solutions – models of agents with infinitely nested beliefs correspond to agent functions which are not computable. It is natural to consider finite nestings. We follow [1, 8, 9], extend our previous work [13], and construct nested I-POMDPs bottom-up. Assume a set of physical states of the world S, and two agents, now denoted i and $-i$. Agent i's 0-th level beliefs, $b_{i,0}$, are probability distributions over S. Its 0-th level types, $\Theta_{i,0}$, contain its 0-th level beliefs, and its frames, and analogously for agent $-i$. 0-th level types are, therefore, POMDPs.[5] i's first level beliefs, $b_{i,1}$, are probability distributions over physical

[5] In 0-th level types the other agent's actions are folded into the T and O functions.

states and 0-level types of agent $-i$. i's first level types consists of its first level beliefs and frames. Second level beliefs are defined in terms of first level types and so on. Formally, define spaces:

$$\begin{array}{rclrcl} IS_{i,0} & = & S, & \Theta_{-i,0} & = & \{\langle b_{-i,0}, \widehat{\theta}_{-i}\rangle : b_{-i,0} \in \Delta(IS_{-i,0})\}, \\ IS_{i,1} & = & IS_{i,0} \times \Theta_{-i,0}, & \Theta_{-i,1} & = & \{\langle b_{-i,1}, \widehat{\theta}_{-i}\rangle : b_{-i,1} \in \Delta(IS_{-i,1})\}, \\ \cdot & & & & \cdot & \\ \cdot & & & & \cdot & \\ \cdot & & & & \cdot & \\ IS_{i,l} & = & IS_{i,l-1} \times \Theta_{-i,l-1}, & \Theta_{-i,l} & = & \{\langle b_{-i,l}, \widehat{\theta}_{-i}\rangle : b_{-i,l} \in \Delta(IS_{-i,l})\}, \\ \cdot & & & & \cdot & \\ \cdot & & & & \cdot & \\ \cdot & & & & \cdot & \end{array}$$

Definition 4. (Finitely Nested I-POMDP) *A finitely nested I-POMDP of agent i, I-POMDP$_{i,l}$, is:*

$$\text{I-POMDP}_{i,l} = \langle IS_{i,l}, A, T_i, \Omega_i, O_i, R_i\rangle \tag{10}$$

The parameter l will be called the strategy level of the finitely nested I-POMDP. The belief update, value function, and the optimal actions for finitely nested I-POMDPs are computed using Equation 3 and Equation 9, but recursion is guaranteed to terminate with 0-th level models.

Agents which are more strategic are capable of modeling others at deeper levels (i.e., all levels up to their own strategy level l), but are always only boundedly optimal. As such, these agents could fail to predict the strategy of a more sophisticated opponent. The fact that the computability of an agent function implies that the agent may be suboptimal during interactions has been pointed out by Binmore [4], and proved more recently in [25]. Intuitively, the difficulty is that an aget's unbounded optimality has to include the capability to model the other agent's modeling the original agent. This leads to impossibility result due to self-reference, very similar Gödel's incompleteness theorem and the halting problem [7]. On a positive note, convergence results in [19] strongly suggest that approximate optimality is achievable, although their applicability to our work remains open.

5.1 Properties of I-POMDPs

In this section we establish two important properties, namely convergence of value iteration and piece-wise linearity and convexity of the value function for finitely nested I-POMDPs.

5.1.1 Convergence of Value Iteration

For an agent i and its *I-POMDP$_{i,l}$*, we can show that the sequence of value functions $\{U^n\}$ where n is the horizon, obtained by value iteration defined in Eq. 8 , converges to a unique fixed-point, U^*.

Let us define a *backup* operator $H : U \rightarrow U$ such that $U^n = HU^{n-1}$. In order to prove the convergence result, we will first establish the properties of the *backup* operator, H.

Lemma 1 (Isotonicity). *For any finitely nested I-POMDP value functions V and U, if $V \leq U$, then $HV \leq HU$.*

The proof of this lemma is analogous to one given in, say, [15], for POMDPs. It is also sketched in the Appendix. Another important property exhibited by the backup operator is the property of contraction.

Lemma 2 (Contraction). *For any finitely nested I-POMDP value functions V, U and a discount factor $\gamma \in (0, 1)$, $||HV - HU||_\infty \leq \gamma||V - U||_\infty$.*

The proof of this lemma is again similar to the corresponding one in POMDPs which is given in [15]. The proof makes use of Lemma 1. Alternatively, the contraction property can be verified using Blackwell's sufficient conditions for a contraction [30].

Under the contraction property of H, and noting that the space of value functions along with the supremum norm forms a complete normed space (Banach space), we can apply the Contraction Mapping Theorem [30] to show that value iteration for I-POMDPs converges to a unique fixed point (optimal solution). The following theorem captures this result.

Theorem 1 (Convergence). *For any finitely nested I-POMDP, the value iteration algorithm starting from any arbitrary well-defined value function converges to a unique fixed-point.*

The detailed proof of this theorem is included in the Appendix.

5.1.2 Piecewise Linearity and Convexity

Another property that carries over from POMDPs to finitely nested I-POMDPs is the piecewise linearity and convexity (PWLC) of the value function. Establishing this property allows us to decompose the I-POMDP value function into a set of *alpha* vectors, each of which represents a policy tree. The PWLC property enables us to work with sets of alpha vectors rather than perform value iteration over the continuum of agent types. Theorem 2 below states the PWLC property of the I-POMDP value function.

Theorem 2 (PWLC). *For any finitely nested I-POMDP, U is piecewise linear and convex.*

The complete proof of Theorem 2 is available for reference in the Appendix. The proof is similar to Smallwood and Sondik's proof [29] for POMDPs and proceeds by induction. The basis case is established by considering the horizon 1 value function. Showing the PWLC for the inductive step requires substituting the belief update (Eq. 3) into Eq. 8, followed by factoring out the belief from both terms of the equation.

6 Interactive Influence Diagrams

A useful method of representing POMDPs and solving them on-line is based on influence diagrams (IDs). Solving an ID for a particular value of agent's initial belief results in a conditional plan. This plan is a special case of solving a POMDP off-line for all possible belief states. We extend an analogous graphical representation to I-POMDPs and define interactive influence diagrams (IIDs). A generic IID is depicted in Fig 1; we omit the temporal dimension here for simplicity.

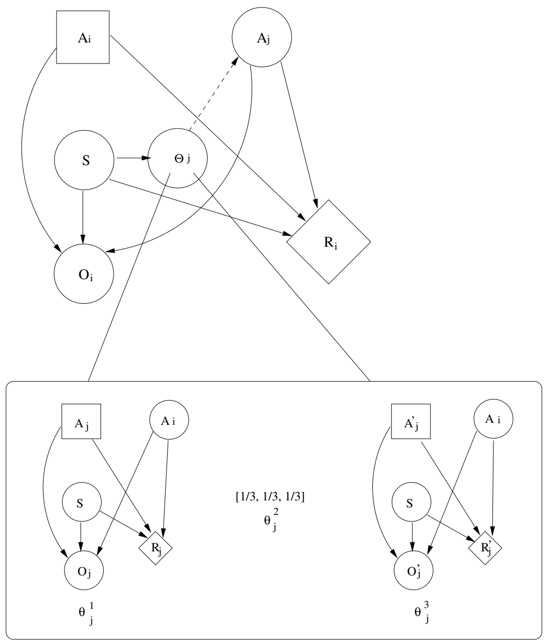

Figure 1: A generic interactive influence diagram.

An interactive influence diagram includes the alternative actions of the other agent, A_j, and a node representing the alternative models of the other agent, Θ_j. The node Θ_j is a complex node, with details of three alternative models that agent i has of agent j shown. In Fig 1 two of these models, θ_j^1, and θ_j^3 are j's possible types. θ_j^2 is a no-information model. In Fig 1 we further assumed that in each of θ_j^1, and θ_j^3 agent j has a no-information model of i. That is, j models i's actions as being equally likely and thus there is no further nesting of models.

Apart from complex nodes, whose probability gets updated in usual Bayesian fashion, IIDs contain a new kind of links, called action links, like the one connecting complex node Θ_j and node A_j. These links indicate that the actions depend on which actions are Bayesian rational for various agent types, as the term $Pr(a_j^{t-1}|\theta_j^{t-1})$ in Eq. 3 specifies.

If an IID contains finitely nested models, with the maximum level of nesting l and the number of alternative types of other agents bound by M, then solving the IID amounts to solving at most M^l classical influence diagrams. Dynamic programming can reduce this number to only $\Theta(Ml)$ influence diagrams. In both of these simple cases solving the IID is NP-hard.

7 Example: Multi-agent Tiger Game

To illustrate optimal sequential behavior of agents in multi-agent settings we apply our I-POMDP framework to the multi-agent tiger game. The traditional tiger game (single agent, defined in [18]) resembles a game-show situation in which the decision maker has to choose to open one of two doors behind which lies either a valuable prize or a dangerous tiger. Apart from actions that open doors, the subject has the option of listening for the tiger's growl. However, the subject's hearing is imperfect, with given percentages of false positive and false negative occurrences. Optimal strategies involve various sequences of listening and opening of doors, depending on the initial belief, reliability of the agent's hearing and the time horizon [18]. In what follows, we introduce a multi-agent version of this tiger game, understand the I-POMDP belief update using a simple setting of the tiger game, compute solutions to the tiger game.

7.1 Definition

We give the problem statement of the multi-agent tiger game below.

Action	State	TL	TR
$< OL, * >$	*	0.5	0.5
$< OR, * >$	*	0.5	0.5
$< *, OL >$	*	0.5	0.5
$< *, OR >$	*	0.5	0.5
$< L, L >$	TL	1.0	0
$< L, L >$	TR	0	1.0

Action/State	TL	TR
$< OR, OR >$	+20	-50
$< OL, OL >$	-50	+20
$< OR, OL >$	-100	-100
$< OL, OR >$	-100	-100
$< L, L >$	-2	-2
$< L, OR >$	+9	-101
$< OR, L >$	+9	-101
$< L, OL >$	-101	+9
$< OL, L >$	-101	+9

Action	State	GL,CL	GL,CR	GL,S	GR,CL	GR,CR	GR,S
L, L	$TL, *$	0.85*0.2	0.85*0.2	0.85*0.6	0.15*0.2	0.15*0.2	0.15*0.6
L, L	$TR, *$	0.15*0.2	0.15*0.2	0.15*0.6	0.85*0.2	0.85*0.2	0.85*0.6
L, OL	$*, *$	0.5*0.6	0.5*0.2	0.5*0.2	0.5*0.6	0.5*0.2	0.5*0.2
L, OR	$*, *$	0.5*0.2	0.5*0.6	0.5*0.2	0.5*0.2	0.5*0.6	0.5*0.2
$OL, *$	$*, *$	1/6	1/6	1/6	1/6	1/6	1/6
$OR, *$	$*, *$	1/6	1/6	1/6	1/6	1/6	1/6

Table 1: Transition, reward, and observation function for the multiagent Tiger game.

Multiagent Tiger Problem The multiagent tiger game resembles a game-show like situation in which there are two doors, behind one is a tiger ready to eat anyone who opens the door, and behind another is a pot of gold. The game is played by two agents,i and j, who are unaware of where the gold is, and each can either open any one of the doors, or listen. The tiger emits a growl periodically, which reveals its position but only with some certainty. Additionally, each agent can also hear a creak with some certainty, if the other agent opens a door. We will assume that neither agent can perceive other's observations nor actions. Now suppose that the two agents must co-ordinate, i.e. each gets twice the reward if both open the correct door (behind which is the pot of gold), and each gets half the penalty if both open the wrong door (behind which is the tiger). The transition, reward, and observation functions describing the game are shown in Table 1. In such a situation, what is the sequence of actions conditioned on observations that each agent should take to maximize its gains over a period of fixed number of steps?

When an agent makes its choice in the multiagent tiger game, it has to consider what it believes about the location of the tiger, as well as whether the other agent will listen or open a door, which in turn depends on the other agent's beliefs. We make a simplifying assumption that agent i knows all of agent j's properties except its belief. This allows us to represent i's uncertainty in terms of the location of the tiger and j's beliefs.

7.2 Example of the Belief Update

In Section 4, we presented the belief update process in I-POMDPs (Eq. 3). As we mentioned before, an agent must continuously update its beliefs, since its optimal action is conditioned on its belief. It may be worthwile to understand the belief update step, before we look at the solutions.

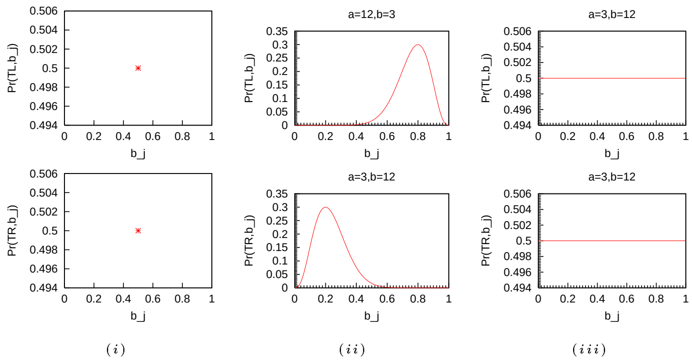

Figure 2: Three examples of level 1 belief states of agent i. In each case i has no information about the tiger's location. (i) Agent i knows that j does not know the location of the tiger, (ii) Agent i believes that j is somewhat informed about the location of the tiger, (iii) Agent i is uninformed about j's beliefs about tiger's location.

When an agent makes its choice in the multiagent tiger game, it has to consider what it believes about the location of the tiger, as well as whether the other agent will listen or open a door, which in turn depends on the other agent's beliefs.

In Fig. 2 we show some example level 1 beliefs of agent i. Fig. $2(i)$ displays a belief state according to which, i knows that j is uninformed about the location of the tiger. In Fig. $2(ii)$ and (iii), we show two level 1 beliefs of agent i using the **Beta** family of distributions. Since agent i's state space is continuous, i's belief is defined using a probability density function (p.d.f.). According to the belief state Fig. $2(ii)$, when the tiger is on the left, i believes that j thinks the tiger is most likely on the left. When the tiger is on the right, i believes that j thinks the tiger is most likely on the right. In other words, i believes that j's beliefs reflect the true situation. Additionally, agent i's belief over the location of the tiger is computed by summing out j's beliefs (marginalization). In Fig. $2(ii)$, i believes with equal probability that the tiger is behind one of the two doors. Fig. $2(iii)$ shows a belief state according to which i is uninformed about j's beliefs.

To make the presentation more transparent we decompose the I-POMDP belief update (Eq. 3 into two steps:

- *Prediction* When an agent, say i in the tiger game, performs a control action a_i^{t-1}, the predicted belief state is,

$$\begin{aligned}\widehat{b}_i^t = Pr(is^t|a_i^{t-1}, a_j^{t-1}, b_i^{t-1}) \quad &= \textstyle\sum_{is^{t-1}|\widehat{\theta}_j^{t-1}=\widehat{\theta}_j^t} b_i^{t-1}(is^{t-1}) Pr(a_j^{t-1}|\theta_j^{t-1}) T(s^{t-1}, a^{t-1}, s^t) \\ &\times \sum_{o_j^t} O_j(s^t, a^{t-1}, o_j^t) \tau_{\theta_j^t}(b_j^{t-1}, a_j^{t-1}, o_j^t, b_j^t)\end{aligned} \tag{11}$$

- *Correction* When agent i perceives an observation, o_i^t, the intermediate belief state, $Pr(\cdot|a_i^{t-1}, a_j^{t-1}, b_i^{t-1})$, is corrected according to,

$$b_i^t = Pr(is^t|o_i^t, a_i^{t-1}, b_i^{t-1}) = \alpha \sum_{a_j^{t-1}} O_i(s^t, a^{t-1}, o_i^t) Pr(is^t|a_i^{t-1}, a_j^{t-1}, b_i^{t-1}) \tag{12}$$

where α is the normalizing constant.

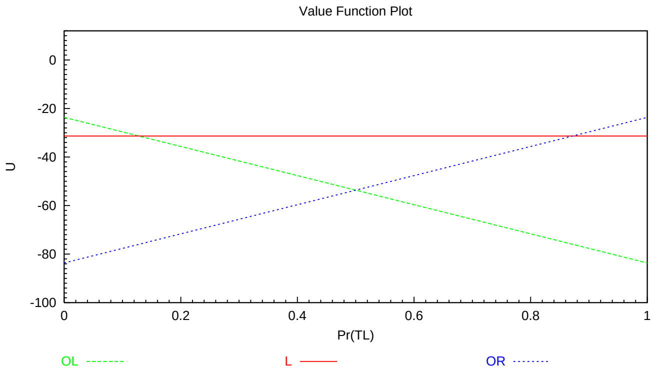

Figure 3: Value function of an agent whose strategy level is 0. We assumed that the other agent's actions are equally likely. Note that listening is the best action if the agent does not have any information about the tiger's location.

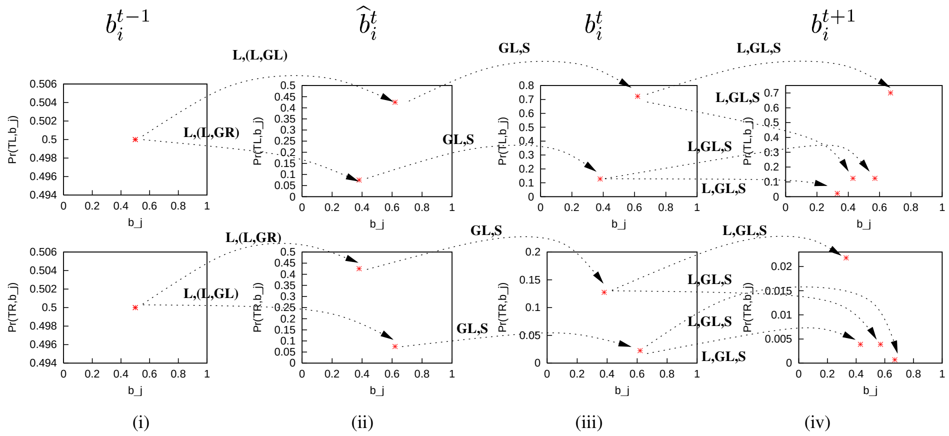

Figure 4: A trace of the belief update of agent i upon its listening (L) and observing tiger's growl on the left and no creaks (GL,S), followed by another listen action and the same observations. (i) Prior belief of agent i. (ii) i's belief after the prediction step given that both agents perform the listening actions. Agent i knows j will listen because i knows j has no information regarding the location of the tiger (see Fig. 3). The single belief point of j is updated given its two possible observations – hearing a growl from left or right (GL or GR). (iii) Posterior belief after the correction step in which i observes GL,S. The marginal for the tiger being on the left is now greater than the marginal for the tiger being on the right. (iv) Posterior belief after one more step of the belief update, during which i listens and again observes GL,S.

To illustrate the belief update, let us trace the update process using the multiagent tiger game. We first note that for a setting in which two agents are playing the tiger game with agent i modeling the other as a no-information model, i opens the left door if its belief of the tiger being on the left is less than 0.128, listens if its belief is between 0.128 and 0.872, and opens the right door if its belief is greater than 0.872. The value function that prescribes this policy is shown in Fig. 3.

In Fig. 4, we display the example trace through the belief update that starts with a prior belief, and then assumes that agent i listens and hears a growl from the left and no creaks. Thereafter the agent again listens and gets the same observations. We will assume, as before, that agent i thinks that j models i using a no-information model. In the first column of Fig. 4, we display agent i's prior belief – i knows that j is uninformed of the location of the tiger. Let us assume that i listens and hears a growl from the left and no creaks. The second column of Fig. 4 displays the *predicted* belief after i performs the listen action(Eq. 11).

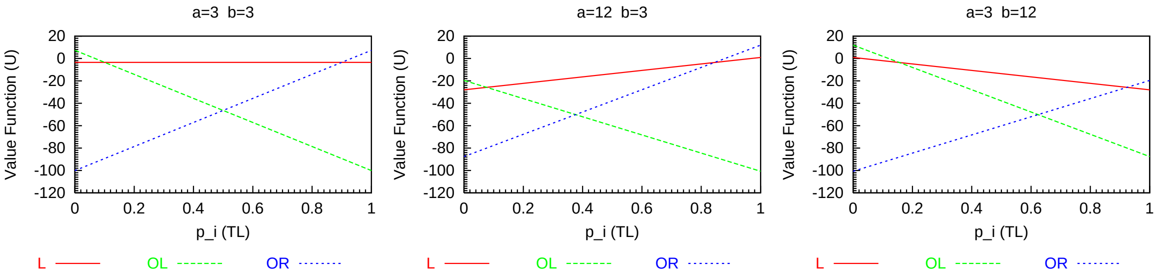

Figure 5: Value functions in two-agent tiger game. The probability of the second agent assigning a probability to tiger being behind the left door is assumed to be beta with parameters: (a) $a = b = 3$, (b) $a = 12$ and $b = 3$, (c) $a = 3$ and $b = 12$. Horizon is 1.

As part of the prediction step, agent i must solve j's model to obtain j's optimal action when its belief is 0.5 (term $Pr(a_j^{t-1}|\theta_j)$ in Eq. 11). Referring to the policy described previously, this action turns out to be listen. The prediction step evaluates to zero for all other actions of agent j. Additionally, i also updates j's belief assuming j listens, and hears the tiger growling from either the left or right (term $\tau_{\theta_j^t}(b_j^{t-1}, a_j^{t-1}, o_j^t, b_j^t)$ in Eq. 11). Agent j's updated beliefs are 0.6167 and 0.3833, respectively. If the tiger is on the left, since j is more likely to hear a growl from left, agent i believes j's former belief to be more likely. When the tiger is on the right, since j is more likely to hear a growl from right, i believes j's latter belief to be more likely. The third column of Fig. 4 shows the posterior belief after the *correction* step. The belief in column (ii) is updated to this belief once i hears a growl from the left and no creaks. The resulting marginal probability of the tiger being on the left is much more than the marginal of the tiger being on the right. If we now assume that i again listens and hears the tiger growling from the left and no creaks, the fourth column of Fig. 4 results. The optimal action for i is to open the right door.

7.3 Example Solutions

An agent's choice in the multi-agent version is guided by what it believes about the location of the tiger as well as the other agent's action, which in turn depends on the other agent's beliefs. For simplicity, let us assume that the first agent, i, models the second agent, j, using intentional models, each of which contains a no-information model of the first agent, i. Thus, i assumes that j models i's three actions as equally likely. Of course j also may assign some probability to the location of the tiger, which is its belief $b_j = p_j(TL)$. It turns out that j's optimal policy dictates that j should open doors if his certainty level about tiger location reaches 0.128. That is, j should open the left door if $p_j(TL) \leq 0.128$ and open the right door if $p_j(TL) \geq 0.872$.

Agent i's belief is over the location of the tiger and models of j, each of which differs only in j's belief, i.e., $p_j(TL)$. Let us assume that i's belief as to the value of $p_j(TL)$ is described by a *beta* distribution with parameters a and b. If $a = b = 3$, the distribution peaks around the value of 0.5, which indicates that i believes that j is likely to not know the tiger's location. When, for example, $a = 12$ and $b = 3$, beta gives more likelihood to high values of $p_j(TL)$, with most likely value being about 0.8. Agent i's value functions for different values of a and b for horizon 1 are shown in Fig. 5.

The first plot in Fig. 5 is symmetric since i's belief as to j's belief about the location of the tiger is symmetric. The second plot depicts the value function for $a = 12$ and $b = 3$. It gives preference to opening the right door. This is because i thinks that j likely believes that tiger is behind the left door, thus j will tend to open the right door and coordinated action is rewarded. In the third plot, for $a = 3$ and $b = 12$, the value function indicates that beliefs that tiger is on the right is more valuable. There is preference for opening the left door, since i believes that j most likely believes that the tiger is behind the right door.

Finally, Fig. 6 is an example of the value function for I-POMDP for $a = 12$ and $b = 3$ and for time horizon equal to 2. Again, beliefs indicating that tiger is on the left are more valuable since i thinks that j likely believes that tiger is on the left. Let us note that the value functions in all cases have the PWLC property, and that for time horizon $T = 2$ values refine the values for the single step game. This illustrates the general results obtained in Section 4.3, of PWLC of value function and convergence of value iteration for I-POMDPs. The corresponding policy graph is also shown. As usual, it prescribes the optimal first action depending on

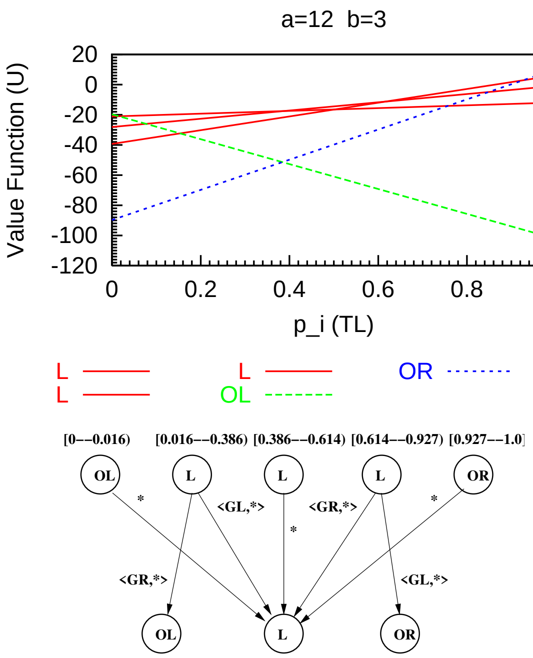

Figure 6: A value function and the corresponding policy graph in the two-agent tiger game. The probability of the second agent assigning a probability to tiger being behind the left door is assumed to be beta with parameters $a = 12$, $b = 3$. Horizon is 2.

the initial belief as to the tiger's location. The second action is then conditioned on the observation obtained after the first action.

8 Conclusions

This paper proposes a general multi-agent version of partially observable Markov decision processes, called interactive POMDPs. We presented assumptions, some basic properties and solution method. Analogously to POMDPs, I-POMDPs provide both reactive as well as deliberative architectures for controlling agent behavior.

One of the avenues of future research centers on proving further formal properties of I-POMDPs. Another concentrates on developing efficient approximation techniques for solving I-POMDPs. As for POMDPs, development of approximate approaches to I-POMDPs is crucial for moving beyond toy problems.

A Appendix

In this section, we give proofs of several properties of I-POMDPs, stated earlier. All proofs that follow assume that the I-POMDP value function is well-defined. One method of ensuring that the value function is well-defined is by restricting the nesting of beliefs in an agent's type to a finite depth.

Equation 8, which defines value iteration in I-POMDPs can be rewritten in the following form, $U^n = HU^{n-1}$. Here, $H : U \to U$ is a *backup* operator, and is defined as,

$$HU^{n-1}(\theta_i) = \max_{a_i \in A_i} h(\theta_i, a_i, U^{n-1})$$

where $h : \Theta_i \times A_i \times U \to \mathbb{R}$ is,

$$h(\theta_i, a_i, U) \quad = \sum_{is} b_i(is) ER_i^{a_j}(is, a_i) + \gamma \sum_{o \in \Omega_i} Pr(o_i|a_i, b_i) U(\langle \hat{\theta}_i, SE_{\theta_i}(b_i, a_i, o_i) \rangle)$$

Lemmas 1 and 2 establish important properties of the backup operator. Proof of Lemma 1 is given below, and proof of Lemma 2 follows thereafter.

Proof of Lemma 1. Select arbitrary well-defined value functions V and U such that $V(\theta_i) \leq U(\theta_i) \;\; \forall \theta_i \in \Theta_i$. Let θ_i be an arbitrary type of agent i.

$$\begin{aligned}
HV(\theta_i) &= \max_{a_i \in A_i} \textstyle\sum_{is} b_i(is) ER_i^{a_j}(is, a_i) + \gamma \sum_{o \in \Omega_i} Pr(o_i|a_i, b_i) V(\langle \hat{\theta}_i, SE_{\theta_i}(b_i, a_i, o_i) \rangle) \\
&= \textstyle\sum_{is} b_i(is) ER_i^{a_j}(is, a_i^*) + \gamma \sum_{o \in \Omega_i} Pr(o_i|a_i^*, b_i) V(\langle \hat{\theta}_i, SE_{\theta_i}(b_i, a_i^*, o_i) \rangle) \\
&\leq \textstyle\sum_{is} b_i(is) ER_i^{a_j}(is, a_i^*) + \gamma \sum_{o \in \Omega_i} Pr(o_i|a_i^*, b_i) U(\langle \hat{\theta}_i, SE_{\theta_i}(b_i, a_i^*, o_i) \rangle) \\
&\leq \max_{a_i \in A_i} \textstyle\sum_{is} b_i(is) ER_i^{a_j}(is, a_i^*) + \gamma \sum_{o \in \Omega_i} Pr(o_i|a_i^*, b_i) U(\langle \hat{\theta}_i, SE_{\theta_i}(b_i, a_i^*, o_i) \rangle) \\
&= HU(\theta_i)
\end{aligned}$$

Since θ_i is arbitrary, $HV \leq HU$. □

Proof of Lemma 2. Assume two arbitrary well defined value functions V and U such that $V \leq U$. From Lemma 1 it follows that $HV \leq HU$. Let θ_i be an arbitrary type of agent i.

$$\begin{aligned}
0 \; &\leq HU(\theta_i) - HV(\theta_i) \\
&= \max_{a_i \in A_i} \textstyle\sum_{is} b_i(is) ER_i^{a_j}(is, a_i) + \gamma \sum_{o \in \Omega_i} Pr(o_i|a_i, b_i) U(\langle \hat{\theta}_i, SE_{\theta_i}(b_i, a_i, o_i) \rangle) - \\
&\quad \max_{a_i \in A_i} \textstyle\sum_{is} b_i(is) ER_i^{a_j}(is, a_i) + \gamma \sum_{o \in \Omega_i} Pr(o_i|a_i, b_i) V(\langle \hat{\theta}_i, SE_{\theta_i}(b_i, a_i, o_i) \rangle) \\
&\leq \textstyle\sum_{is} b_i(is) ER_i^{a_j}(is, a_i^*) + \gamma \sum_{o \in \Omega_i} Pr(o_i|a_i^*, b_i) U(\langle \hat{\theta}_i, SE_{\theta_i}(b_i, a_i^*, o_i) \rangle) - \\
&\quad \textstyle\sum_{is} b_i(is) ER_i^{a_j}(is, a_i^*) - \gamma \sum_{o \in \Omega_i} Pr(o_i|a_i^*, b_i) V(\langle \hat{\theta}_i, SE_{\theta_i}(b_i, a_i^*, o_i) \rangle) \\
&= \gamma \textstyle\sum_{o \in \Omega_i} Pr(o_i|a_i^*, b_i) U(\langle \hat{\theta}_i, SE_{\theta_i}(b_i, a_i^*, o_i) \rangle) - \gamma \sum_{o \in \Omega_i} Pr(o_i|a_i^*, b_i) V(\langle \hat{\theta}_i, SE_{\theta_i}(b_i, a_i^*, o_i) \rangle) \\
&= \gamma \textstyle\sum_{o \in \Omega_i} Pr(o_i|a_i^*, b_i) \Big[U(\langle \hat{\theta}_i, SE_{\theta_i}(b_i, a_i^*, o_i) \rangle) - V(\langle \hat{\theta}_i, SE_{\theta_i}(b_i, a_i^*, o_i) \rangle) \Big] \\
&\leq \gamma \textstyle\sum_{o \in \Omega_i} Pr(o_i|a_i^*, b_i) ||U - V||_\infty \\
&= \gamma ||U - V||_\infty
\end{aligned}$$

As the supremum norm is symmetrical, a similar result can be derived for $HV(\theta_i) - HU(\theta_i)$. Since θ_i is arbitrary, the Contraction property follows, i.e. $||HV - HU||_\infty \leq ||V - U||_\infty$. □

Lemmas 1 and 2 provide the stepping stones for proving Theorem 1. Proof of Theorem 1 follows from a straightforward application of the Contraction Mapping Theorem. We state the Contraction Mapping Theorem below:

Theorem 3 (Contraction Mapping Theorem). *If (S, ρ) is a complete metric space and $T : S \rightarrow S$ is a contraction mapping with modulus γ, then*

1. *T has exactly one fixed point U^* in S, and*
2. *The sequence $\{U^n\}$ converges to U^*.*

Proof of Theorem 1 follows.

Proof of Theorem 1. It is not difficult to see that the normed space $(U, ||\cdot||_\infty)$ is complete w.r.t the metric induced by the supremum norm. Lemma 2 establishes the contraction property of the backup operator, H. Using Theorem 3, and substituting T with H, convergence of value iteration in I-POMDPs to a unique fixed point is established. □

Let us turn our attention now to proving another interesting property of the value function which is its piecewise linearity and convexity (PWLC). Our proof of PWLC (Theorem 2) is inspired in part by the proof in [xxx] of the same property in POMDPs.

Let $\alpha : IS \rightarrow \mathbb{R}$ be a Borel-measurable function. Let the space of such real-valued bounded functions be $B(IS)$. We will now define an inner product.

Definition 5 (Inner product). *Define the inner product,* $\langle \cdot, \cdot \rangle : B(IS) \times \Delta(IS) \rightarrow \mathbb{R}$*, by*

$$\langle \alpha, b_i \rangle = \sum_{is} b_i(is)\alpha(is)$$

The next lemma establishes the bilinearity of the inner product defined above.

Lemma 3 (Bilinearity). *For any* $s, t \in \mathbb{R}$*,* $f, g \in B(IS)$*, and* $b, \lambda \in \Delta(IS)$ *the following equalities hold:*

$$\langle sf + tg, b \rangle = s\langle f, b \rangle + t\langle g, b \rangle$$
$$\langle f, sb + t\lambda \rangle = s\langle f, b \rangle + t\langle f, \lambda \rangle$$

We are now ready to give the proof of Theorem 2. Theorem 4 restates Theorem 2 mathematically, and its proof follows thereafter.

Theorem 4 (PWLC). *The value function,* U^n*, in I-POMDP is* piece-wise linear and convex *(PWLC). Mathematically,*

$$U^n(\theta_i) = \max_{\alpha^n \in \Gamma^n} \sum_{is} b_i(is)\alpha^n(is) \qquad n = 1, 2, ...$$

Proof of Theorem 4. **Basis Step:** $n = 1$

From Bellman's Dynamic Programming equation,

$$U^1(\theta_i) = \max_{a_i} \sum_{is} b_i(is)ER^{a_j}(is, a_i) \tag{13}$$

where $ER^{a_j}(is, a_i) = \sum_{a_j} R(is, a_i, a_j)Pr(a_j|is)$.

Here, $ER^{a_j}(\cdot)$ represents the expectation of R w.r.t. agent j's actions. Eq. 13 represents an inner product and using Lemma 3, the inner product is linear in b_i. By selecting the maximum of a set of linear vectors, we obtain a PWLC horizon 1 value function.

Inductive Hypothesis: Suppose that $U^{n-1}(\theta_i)$ is PWLC. Mathematically,

$$U^{n-1}(\theta_i) = \max_{\alpha^{n-1} \in \Gamma^{n-1}} \langle \alpha^{n-1}, b_i^t \rangle = \max_{\alpha^{n-1} \in \Gamma^{n-1}} \sum_{is^t} b_i^t(is^t)\alpha^{n-1}(is^t) \tag{14}$$

Inductive Proof: To show that $U^n(\theta_i)$ is PWLC.

$$U^n(\theta_i) = \max_{a_i^{t-1}} \left\{ \sum_{is^{t-1}} b_i^{t-1}(is^{t-1})ER^{a_j}(is^{t-1}, a_i^{t-1}) + \gamma \sum_{o_i^t} Pr(o_i^t|a_i^{t-1}, b_i^{t-1})U^{n-1}(\theta_i) \right\}$$

¿From the inductive hypothesis,

$$U^n(\theta_i) = \max_{a_i^{t-1}} \left\{ \sum_{is^{t-1}} b_i^{t-1}(is^{t-1})ER^{a_j}(is^{t-1}, a_i^{t-1}) + \gamma \sum_{o_i^t} Pr(o_i^t|a_i^{t-1}, b_i^{t-1}) \max_{\alpha^{n-1} \in \Gamma^{n-1}} \sum_{is^t} b_i^t(is^t)\alpha^{n-1}(is^t) \right\}$$

Let $l(b_i^{t-1}, a_i^{t-1}, o_i^t)$ be the index of the alpha vector that maximizes the integral at $b_i^t = SE(b_i^{t-1}, a_i^{t-1}, o_i^t)$. Then,

$$U^n(\theta_i) = \max_{a_i^{t-1}} \left\{ \sum_{is^{t-1}} b_i^{t-1}(is^{t-1})ER^{a_j}(is^{t-1}, a_i^{t-1}) + \gamma \sum_{o_i^t} Pr(o_i^t|a_i^{t-1}, b_i^{t-1}) \sum_{is^t} b_i^t(is^t)\alpha^{n-1}_{l(b_i^{t-1}, a_i^{t-1}, o_i^t)} \right\}$$

Substituting b_i^t with the belief update in Eq. 3, we get,

$$\begin{aligned}
U^n(\theta_i) &= \max_{a_i^{t-1}} \Bigg\{ \sum_{is^{t-1}} b_i^{t-1}(is^{t-1}) ER^{a_j}(is^{t-1}, a_i^{t-1}) + \gamma \sum_{o_i^t} Pr(o_i^t | a_i^{t-1}, b_i^{t-1}) \\
&\times \sum_{is^t} \beta \sum_{is^{t-1}} b_i^{t-1}(is^{t-1}) \Bigg\{ \sum_{a_j^{t-1}} Pr(a_j^{t-1}|m_j^{t-1}) \Bigg[O_i(is^t, a^{t-1}, o_i^t) \sum_{o_j^t} O_j^t(is_j^t, a^{t-1}, o_j^t) \\
&\times \Bigg\{ \tau_{\theta_j^t}(b_j^{t-1}, a_j^{t-1}, o_j^t, b_j^t) T_i(is^{t-1}, a^{t-1}, is^t) \Bigg\} \Bigg] \Bigg\} \alpha^{n-1}_{l(b_i^{t-1}, a_i^{t-1}, o_i^t)}(is^t) \Bigg\} \\
&= \max_{a_i^{t-1}} \Bigg\{ \sum_{is^{t-1}} b_i^{t-1}(is^{t-1}) ER^{a_j}(is^{t-1}, a_i^{t-1}) + \gamma \sum_{is^t} \sum_{is^{t-1}} \sum_{o_i^t} \Bigg\{ b_i^{t-1}(is^{t-1}) \sum_{a_j^{t-1}} Pr(a_j^{t-1}|m_j^{t-1}) \Bigg[O_i(is^t, a^{t-1}, \\
&\times \sum_{o_j^t} O_j^t(is_j^t, a^{t-1}, o_j^t) \Bigg\{ \tau_{\theta_j^t}(b_j^{t-1}, a_j^{t-1}, o_j^t, b_j^t) T_i(is^{t-1}, a^{t-1}, is^t) \Bigg\} \Bigg] \Bigg\} \alpha^{n-1}_{l(b_i^{t-1}, a_i^{t-1}, o_i^t)}(is^t) \Bigg\}
\end{aligned}$$

Rearranging the terms of the equation,

$$\begin{aligned}
U^n(\theta_i) &= \max_{a_i^{t-1}} \Bigg\{ \sum_{is^{t-1}} b_i^{t-1}(is^{t-1}) \Bigg\{ ER^{a_j}(is^{t-1}, a_i^{t-1}) + \gamma \sum_{is^t} \sum_{o_i^t} \Bigg\{ \sum_{a_j^{t-1}} Pr(a_j^{t-1}|m_j^{t-1}) \Bigg[O_i(is^t, a^{t-1}, o_i^t) \\
&\times \sum_{o_j^t} O_j^t(is_j^t, a^{t-1}, o_j^t) \Bigg\{ \tau_{\theta_j^t}(b_j^{t-1}, a_j^{t-1}, o_j^t, b_j^t) T_i(is^{t-1}, a^{t-1}, is^t) \Bigg\} \Bigg] \Bigg\} \alpha^{n-1}_{l(b_i^{t-1}, a_i^{t-1}, o_i^t)}(is^t) \Bigg\} \Bigg\} \\
&= \max_{a_i^{t-1}} \sum_{is^{t-1}} b_i^{t-1}(is^{t-1}) \alpha^n_{a_i}(is^{t-1})
\end{aligned}$$

Therefore,

$$U^n(\theta_i) = \max_{\alpha^n_{a_i} \in \Gamma^n} \sum_{is^{t-1}} b_i^{t-1}(is^{t-1}) \alpha^n_{a_i}(is^{t-1}) = \max_{\alpha^n_{a_i} \in \Gamma^n} \langle \alpha^n_{a_i}, b_i^{t-1} \rangle \tag{15}$$

where,

$$\begin{aligned}
\alpha^n_{a_i^{t-1}}(is^{t-1}) &= ER^{a_j}(is^{t-1}, a_i^{t-1}) + \gamma \sum_{o_i^t} \sum_{is^t} \Bigg\{ \sum_{a_j^{t-1}} Pr(a_j^{t-1}|m_j^{t-1}) \Bigg[O_i(is^t, a^{t-1}, o_i^t) \\
&\times \sum_{o_j^t} O_j^t(is_j^t, a^{t-1}, o_j^t) \Bigg\{ \tau_{\theta_j^t}(b_j^{t-1}, a_j^{t-1}, o_j^t, b_j^t) T_i(is^{t-1}, a^{t-1}, is^t) \Bigg\} \Bigg] \Bigg\} \alpha^{n-1}_{l(b_i^{t-1}, a_i^{t-1}, o_i^t)}(is^t)
\end{aligned}$$

corresponds to a linear vector for an action a_i.

Eq. 15 is an **inner product** and using Lemma 3, it is linear in b_i^{t-1}. Furthermore, maximizing over a set of linear vectors produces a piecewise linear and convex value function. □

References

[1] R. J. Aumann. Interactive epistemology i: Knowledge. *International Journal of Game Theory*, (28):263–300, 1999.

[2] R. J. Aumann and A. Heifetz. Incomplete information. In R. Aumann and S. Hart, editors, *Handbook of Game Theory with Economic Applications, Volume III, Chapter 43*. Elsevier, 2002.

[3] D. S. Bernstein, R. Givan, N. Immerman, and S. Zilberstein. The complexity of decentralized control of markov decision processes. *To appear in Mathematics of Operations Research*, 2002.

[4] K. Binmore. *Essays on Foundations of Game Theory*. Blackwell, 1990.

[5] C. Boutilier. Sequential optimality and coordination in multiagent systems. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, pages 478–485, Aug. 1999.

[6] C. Boutilier, T. Dean, and S. Hanks. Decision-theoretic planning: Structural assumptions and computational leverage. *Journal of Artificial intelligence Research*, 11:1–94, 1999.

[7] A. Brandenburger. The power of paradox: Some recent developments in interactive epistemology. Technical report, Stern School of Business, New York University, http://pages.stern.nyu.edu/ abranden/, Apr. 2002.

[8] A. Brandenburger and E. Dekel. Hierarchies of beliefs and common knowledge. *Journal of Economic Theory*, 59:189–198, 1993.

[9] R. R. Fagin, J. Geanakoplos, J. Y. Halpern, and M. Y. Vardi. A hierarchical approach to modeling knowledge and common knowledge. *International Journal of Game Theory*, (28):331–365, 1999.

[10] R. R. Fagin, J. Y. Halpern, Y. Moses, and M. Y. Vardi. *Reasoning About Knowledge*. MIT Press, 1995.

[11] D. Fudenberg and D. K. Levine. *The Theory of Learning in Games*. MIT Press, 1998.

[12] D. Fudenberg and J. Tirole. *Game Theory*. MIT Press, 1991.

[13] P. J. Gmytrasiewicz and E. H. Durfee. Rational coordination in multi-agent environments. *Autonomous Agents and Multiagent Systems Journal*, 3(4):319–350, 2000.

[14] J. C. Harsanyi. Games with incomplete information played by 'Bayesian' players. *Management Science*, 14(3):159–182, Nov. 1967.

[15] M. Hauskrecht. Value-function approximations for partially observable markov decision processes. *Journal of Artificial Intelligence Research*, (13):33–94, July 2000.

[16] J. Hu and M. P. Wellman. Multiagent reinforcement learning: Theoretical framework and an algorithm. In *Fifteenth International Conference on Machine Learning*, pages 242–250, 1998.

[17] J. B. Kadane and P. D. Larkey. Subjective probability and the theory of games. *Management Science*, 28(2):113–120, Feb. 1982.

[18] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, (2), 1998.

[19] E. Kalai and E. Lehrer. Rational learning leads to nash equilibrium. *Econometrica*, (61):1231–1240, 1993.

[20] D. Koller and B. Milch. Multi-agent influence diagrams for representing and solving games. In *Seventeenth International Joint Conference on Artificial Intelligence*, pages 1027–1034, Seattle, Washington, Aug. 2001.

[21] M. L. Littman. Markov games as a framework for multi-agent reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, 1994.

[22] J.-F. Mertens and S. Zamir. Formulation of Bayesian analysis for games with incomplete information. *International Journal of Game Theory*, 14:1–29, 1985.

[23] G. E. Monahan. A survey of partially observable markov decision processes: Theory, models, and algorithms. *Management Science*, pages 1–16, 1982.

[24] R. B. Myerson. *Game Theory: Analysis of Conflict*. Harvard University Press, 1991.

[25] J. H. Nachbar and W. R. Zame. Non-computable strategies and discounted repeated games. *Economic Theory*, 8:103–122, 1996.

[26] J. M. Ooi and G. W. Wornell. Decentralized control of a multiple access broadcast channel. In *Proceedings of the 35th Conference on Decision and Control*, 1996.

[27] M. T. Ranjit Nair, David Pynadath and S. Marsella. Towards computing optimal policies for decentralized pomdps. In *Notes of the 2002 AAAI Workshop on Game Theoretic and Decision Theoretic Agents*, 2002.

[28] S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach (Second Edition)*. Prentice Hall, 2003.

[29] R. D. Smallwood and E. J. Sondik. The optimal control of partially observable markov decision processes over a finite horizon. *Operations Research*, (21):1071–1088, 1973.

[30] N. L. Stokey and L. R. E. *Recursive Methods in Economic Dynamics*. Harvard Univ. Press, 1989.