Interactive POMDP Lite: Towards Practical Planning to Predict and Exploit Intentions for Interacting with Self-Interested Agents

Trong Nghia Hoang and **Kian Hsiang Low**
Department of Computer Science, National University of Singapore
Republic of Singapore
{nghiaht, lowkh}@comp.nus.edu.sg

Abstract

A key challenge in non-cooperative multi-agent systems is that of developing efficient planning algorithms for intelligent agents to interact and perform effectively among boundedly rational, self-interested agents (e.g., humans). The practicality of existing works addressing this challenge is being undermined due to either the restrictive assumptions of the other agents' behavior, the failure in accounting for their rationality, or the prohibitively expensive cost of modeling and predicting their intentions. To boost the practicality of research in this field, we investigate how intention prediction can be efficiently exploited and made practical in planning, thereby leading to efficient intention-aware planning frameworks capable of predicting the intentions of other agents and acting optimally with respect to their predicted intentions. We show that the performance losses incurred by the resulting planning policies are linearly bounded by the error of intention prediction. Empirical evaluations through a series of stochastic games demonstrate that our policies can achieve better and more robust performance than the state-of-the-art algorithms.

1 Introduction

A fundamental challenge in non-cooperative *multi-agent systems* (MAS) is that of designing intelligent agents that can efficiently plan their actions under uncertainty to interact and perform effectively among boundedly rational[1], self-interested agents (e.g., humans). Such a challenge is posed by many real-world applications [Hansen *et al.*, 2004], which include automated electronic trading markets where software agents interact, and traffic intersections where autonomous cars have to negotiate with human-driven vehicles to cross them, among others. These applications can be modeled as partially observable stochastic games (POSGs) in which the agents are self-interested (i.e., non-cooperative) and do not necessarily share the same goal, thus invalidating the use of planning algorithms developed for coordinating cooperative agents (i.e., solving POSGs with common payoffs) [Nair and Tambe, 2003; Seuken and Zilberstein, 2007; Spaan *et al.*, 2011]. Existing planning frameworks for non-cooperative MAS can be generally classified into:

Game-theoretic frameworks. Based on the well-founded classical game theory, these multi-agent planning frameworks [Hansen *et al.*, 2004; Hu and Wellman, 1998][2] characterize the agents' interactions in a POSG using solution concepts such as Nash equilibrium. Such frameworks suffer from the following drawbacks: (a) Multiple equilibria may exist, (b) only the optimal actions corresponding to the equilibria are specified, and (c) they assume that the agents do not collaborate to beneficially deviate from the equilibrium (i.e., no coalition), which is often violated by human agents.

Decision-theoretic frameworks. Unafflicted by the drawbacks of game-theoretic approaches, they extend single-agent decision-theoretic planning frameworks such as the *Markov decision process* (MDP) and *partially observable Markov decision process* (POMDP) to further characterize interactions with the other self-interested agents in a POSG. In particular, the *interactive POMDP* (I-POMDP) framework [Doshi and Gmytrasiewicz, 2009; Doshi and Perez, 2008; Gmytrasiewicz and Doshi, 2005; Rathnasabapathy *et al.*, 2006] is proposed to explicitly account for the bounded rationality of self-interested agents: It replaces POMDP's flat beliefs over the physical states with interactive beliefs over both the physical states and the other agent's beliefs. Empowered by such an enriched, highly expressive belief space, I-POMDP can explicitly model and predict the other agent's intention (i.e., mixed strategy) under partial observability.

However, solving I-POMDP is prohibitively expensive due to the following computational difficulties [Doshi and Perez, 2008; Gmytrasiewicz and Doshi, 2005]: (a) **Curse of dimensionality** – since I-POMDP's interactive belief is over the joint space of physical states and the other agent's beliefs (termed interactive state space in [Doshi and Perez, 2008]), its dimension can be extremely large and possibly infinite; (b) **curse of history** – similar to POMDP, I-POMDP's policy space grows exponentially with the length of planning horizon; and (c) **curse of nested reasoning** – as I-POMDP

[1] Boundedly rational agents are subject to limited cognition and time in making decisions [Gigerenzer and Selten, 2002].

[2] The learning framework of Hu and Wellman [1998] trivially reduces to planning when the transition model is known a priori.

utilizes a nested structure within our agent's belief space to represent its belief over the other agent's belief and the other agent's belief over our agent's belief and so on, it aggravates the effects of the other two curses [Doshi and Perez, 2008].

To date, a number of approximate I-POMDP techniques [Doshi and Gmytrasiewicz, 2009; Doshi and Perez, 2008; Rathnasabapathy *et al.*, 2006] have been proposed to mitigate some of the above difficulties. Notably, *Interactive Particle Filtering* (I-PF) [Doshi and Gmytrasiewicz, 2009] focused on alleviating the curse of dimensionality by generalizing the particle filtering technique to accommodate the multi-agent setting while *Interactive Point-based Value Iteration* [Doshi and Perez, 2008] (I-PBVI) aimed at relieving the curse of history by generalizing the well-known point-based value iteration (PBVI) [Pineau *et al.*, 2003] to operate in the interactive belief space. Unfortunately, I-PF fails to address the curse of history and it is not clear how PBVI or other sampling-based algorithm can be modified to work with a particle representation of interactive beliefs, whereas I-PBVI suffers from the curse of dimensionality because its dimension of interactive belief grows exponentially with the length of planning horizon of the other agent (Section 3). Using interactive beliefs, it is therefore not known whether it is even possible to jointly lift both curses, for example, by extending I-PF or I-PBVI. Furthermore, they do not explicitly account for the curse of nested reasoning. As a result, their use has been restricted to small, simple problems [Ng *et al.*, 2010] (e.g., multiagent Tiger [Gmytrasiewicz and Doshi, 2005; Nair and Tambe, 2003; Doshi and Perez, 2008]).

To tractably solve larger problems, existing approximate I-POMDP techniques such as I-PF and I-PBVI have to significantly reduce the quality of approximation and impose restrictive assumptions (Section 3), or risk not producing a policy at all with the available memory of modern-day computers. This naturally raises the concern of whether the resulting policy can still perform well or not under different partially observable environments, as investigated in Section 5. Since such drastic compromises in solution quality are necessary of approximate I-POMDP techniques to tackle a larger problem directly, it may be worthwhile to instead consider formulating an approximate version of the problem with a less sophisticated structural representation such that it allows an exact or near-optimal solution policy to be more efficiently derived. More importantly, can the induced policy perform robustly against errors in modeling and predicting the other agent's intention? If we are able to formulate such an approximate problem, the resulting policy can potentially perform better than an approximate I-POMDP policy in the original problem while incurring significantly less planning time.

Our work in this paper investigates such an alternative: We first develop a novel intention-aware *nested MDP* framework (Section 2) for planning in fully observable multi-agent environments. Inspired by the cognitive hierarchy model of games [Camerer *et al.*, 2004], nested MDP constitutes a recursive reasoning formalism to predict the other agent's intention and then exploit it to plan our agent's optimal interaction policy. Its formalism is by no means a reduction of I-POMDP. We show that nested MDP incurs linear time in the planning horizon length and reasoning depth. Then, we propose an I-POMDP Lite framework (Section 3) for planning in partially observable multi-agent environments that, in particular, exploits a practical structural assumption: The intention of the other agent is driven by nested MDP, which is demonstrated theoretically to be an effective surrogate of its true intention when the agents have fine sensing and actuation capabilities. This assumption allows the other agent's intention to be predicted efficiently and, consequently, I-POMDP Lite to be solved efficiently in polynomial time, hence lifting the three curses of I-POMDP. As demonstrated empirically, it also improves I-POMDP Lite's robustness in planning performance by overestimating the true sensing capability of the other agent. We provide theoretical performance guarantees of the nested MDP and I-POMDP Lite policies that improve with decreasing error of intention prediction (Section 4). We extensively evaluate our frameworks through experiments involving a series of POSGs that have to be modeled using a significantly larger state space (Section 5).

2 Nested MDP

Given that the environment is fully observable, our proposed nested MDP framework can be used to predict the other agent's strategy and such predictive information is then exploited to plan our agent's optimal interaction policy. Inspired by the cognitive hierarchy model of games [Camerer *et al.*, 2004], it constitutes a well-defined recursive reasoning process that comprises k levels of reasoning. At level 0 of reasoning, our agent simply believes that the other agent chooses actions randomly and computes its best response by solving a conventional MDP that implicitly represents the other agent's actions as stochastic noise in its transition model. At higher reasoning levels $k \geq 1$, our agent plans its optimal strategy by assuming that the other agent's strategy is based only on lower levels $0, 1, \ldots, k-1$ of reasoning. In this section, we will formalize nested MDP and show that our agent's optimal policy at level k can be computed recursively.

Nested MDP Formulation. Formally, nested MDP for agent t at level k of reasoning is defined as a tuple $M_t^k \triangleq \left(S, U, V, T, R, \{\pi_{\text{-}t}^i\}_{i=0}^{k-1}, \phi\right)$ where S is a set of all possible states of the environment; U and V are, respectively, sets of all possible actions available to agents t and -t; $T : S \times U \times V \times S \to [0,1]$ denotes the probability $Pr(s'|s,u,v)$ of going from state $s \in S$ to state $s' \in S$ using agent t's action $u \in U$ and agent -t's action $v \in V$; $R : S \times U \times V \to \mathbb{R}$ is a reward function of agent t; $\pi_{\text{-}t}^i : S \times V \to [0,1]$ is a reasoning model of agent -t at level $i < k$, as defined later in (3); and $\phi \in (0,1)$ is a discount factor.

Nested MDP Planning. The optimal $(h+1)$-step-to-go value function of nested MDP M_t^k at level $k \geq 0$ for agent t satisfies the following Bellman equation:

$$U_t^{k,h+1}(s) \triangleq \max_{u \in U} \sum_{v \in V} \widehat{\pi}_{\text{-}t}^k(s,v)\, Q_t^{k,h+1}(s,u,v)$$
$$Q_t^{k,h+1}(s,u,v) \triangleq R(s,u,v) + \phi \sum_{s' \in S} T(s,u,v,s')\, U_t^{k,h}(s') \quad (1)$$

where the mixed strategy $\widehat{\pi}_{\text{-}t}^k$ of the other agent -t for $k > 0$

is predicted as

$$\widehat{\pi}^k_{\text{-}t}(s,v) \triangleq \begin{cases} \sum_{i=0}^{k-1} p(i)\pi^i_{\text{-}t}(s,v) & \text{if } k>0, \\ |V|^{-1} & \text{otherwise.} \end{cases} \quad (2)$$

where the probability $p(i)$ (i.e., $\sum_{i=0}^{k-1} p(i) = 1$) specifies how likely agent -t will reason at level i; a uniform distribution is assumed when there is no such prior knowledge. Alternatively, one possible direction for future work is to learn $p(i)$ using multi-agent reinforcement learning techniques such as those described in [Chalkiadakis and Boutilier, 2003; Hoang and Low, 2013a]. At level 0, agent -t's reasoning model $\pi^0_{\text{-}t}$ is induced by solving $M^0_{\text{-}t}$. To obtain agent -t's reasoning models $\{\pi^i_{\text{-}t}\}_{i=1}^{k-1}$ at levels $i = 1,\ldots,k-1$, let $Opt^i_{\text{-}t}(s)$ be the set of agent -t's optimal actions for state s induced by solving its nested MDP $M^i_{\text{-}t}$, which recursively involves building agent t's reasoning models $\{\pi^l_t\}_{l=0}^{i-1}$ at levels $l = 0,1,\ldots,i-1$, by definition. Then,

$$\pi^i_{\text{-}t}(s,v) \triangleq \begin{cases} |Opt^i_{\text{-}t}(s)|^{-1} & \text{if } v \in Opt^i_{\text{-}t}(s), \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

After predicting agent -t's mixed strategy $\widehat{\pi}^k_{\text{-}t}$ (2), agent t's optimal policy (i.e., reasoning model) π^k_t at level k can be induced by solving its corresponding nested MDP M^k_t (1).

Time Complexity. Solving M^k_t involves solving $\{M^i_{\text{-}t}\}_{i=0}^{k-1}$, which, in turn, requires solving $\{M^i_t\}_{i=0}^{k-2}$, and so on. Thus, solving M^k_t requires solving M^i_t ($i = 0,\ldots,k-2$) and $M^i_{\text{-}t}$ ($i = 0,\ldots,k-1$), that is, $\mathcal{O}(k)$ nested MDPs. Given $\widehat{\pi}^k_{\text{-}t}$, the cost of deriving agent t's optimal policy grows linearly with the horizon length h as the backup operation (1) has to be performed h times. In turn, each backup operation incurs $\mathcal{O}\big(|S|^2\big)$ time given that $|U|$ and $|V|$ are constants. Then, given agent -t's profile of reasoning models $\{\pi^i_{\text{-}t}\}_{i=0}^{k-1}$, predicting its mixed strategy $\widehat{\pi}^k_{\text{-}t}(s,v)$ (2) incurs $\mathcal{O}(k)$ time. Therefore, solving agent t's nested MDP M^k_t (1) or inducing its corresponding reasoning model π^k_t incurs $\mathcal{O}\big(kh|S|^2\big)$.

3 Interactive POMDP Lite

To tackle partial observability, it seems obvious to first consider generalizing the recursive reasoning formalism of nested MDP. This approach yields two practical complications: (a) our agent's belief over both the physical states and the other agent's beliefs (i.e., a probability distribution over probability distributions) has to be modeled, and (b) the other agent's mixed strategy has to be predicted for each of its infinitely many possible beliefs. Existing approximate I-POMDP techniques address these respective difficulties by (a) using a finite particle representation like I-PF [Doshi and Gmytrasiewicz, 2009] or (b) constraining the interactive state space IS to $IS' = S \times \text{Reach}(B,h)$ like I-PBVI [Doshi and Perez, 2008] where $\text{Reach}(B,h)$ includes the other agent's beliefs reachable from a finite set B of its candidate initial beliefs over horizon length h.

However, recall from Section 1 that since I-PF suffers from the curse of history, the particle approximation of interactive beliefs has to be made significantly coarse to solve larger problems tractably, thus degrading its planning performance. I-PBVI, on the other hand, is plagued by the curse of dimensionality due to the need of constructing the set $\text{Reach}(B,h)$ whose size grows exponentially with h. As a result, it cannot tractably plan beyond a few look-ahead steps for even the small test problems in Section 5. Furthermore, it imposes a restrictive assumption that the true initial belief of the other agent, which is often not known in practice, needs to be included in B to satisfy the absolute continuity condition of interactive beliefs [Doshi and Perez, 2008] (see Appendix C of Hoang and Low [2013b] for more details). So, I-PBVI may not perform well under practical environmental settings where a long planning horizon is desirable or the other agent's initial belief is not included in B. For I-PF and I-PBVI, the curse of nested reasoning aggravates the effects of other curses.

Since predicting the other agent's intention using approximate I-POMDP techniques is prohibitively expensive, it is practical to consider a computationally cheaper yet credible information source providing its intention such as its nested MDP policy. Intuitively, such a policy describes the intention of the other agent with full observability who believes that our agent has full observability as well. Knowing the other agent's nested MDP policy is especially useful when the agents' sensing and actuation capabilities are expected to be good (i.e., accurate observation and transition models), as demonstrated in the following simple result:

Theorem 1 *Let* $\widehat{Q}^n_{\text{-}t}(s,v) \triangleq |U|^{-1}\sum_{u\in U} Q^{0,n}_{\text{-}t}(s,v,u)$ *and* $\widehat{Q}^n_{\text{-}t}(b,v)$ *denote* n*-step-to-go values of selecting action* $v \in V$ *in state* $s \in S$ *and belief* b*, respectively, for the other agent* -t *using nested MDP and I-POMDP at reasoning level* 0 *(i.e., MDP and POMDP). If* $b(s) \geq 1-\epsilon$ *and*

$$\forall\,(s,v,u)\;\exists\,(s',o)\;Pr(s'|s,v,u)\geq 1-\frac{\epsilon}{2} \wedge Pr(o|s,v)\geq 1-\frac{\epsilon}{2}$$

for some $\epsilon \geq 0$*, then*

$$\left|\widehat{Q}^n_{\text{-}t}(s,v) - \widehat{Q}^n_{\text{-}t}(b,v)\right| \leq \epsilon\,\mathcal{O}\left(\frac{R_{\max}-R_{\min}}{1-\phi}\right) \quad (4)$$

where $R_{\max}$ *and* $R_{\min}$ *denote agent* -t*'s maximum and minimum immediate payoffs, respectively.*

Its proof is given in Appendix A of Hoang and Low [2013b]. Following from Theorem 1, we conjecture that, as ϵ decreases (i.e., observation and transition models become more accurate), the nested MDP policy $\pi^0_{\text{-}t}$ of the other agent is more likely to approximate the exact I-POMDP policy closely. Hence, the nested MDP policy serves as an effective surrogate of the exact I-POMDP policy (i.e., true intention) of the other agent if the agents have fine sensing and actuation capabilities; such a condition often holds for typical real-world environments.

Motivated by the above conjecture and Theorem 1, we propose an alternative I-POMDP Lite framework by exploiting the following structural assumption: The intention of the other agent is driven by nested MDP. This assumption allows the other agent's intention to be predicted efficiently by computing its nested MDP policy, thus lifting I-POMDP's curse of nested reasoning (Section 2). More importantly, it enables both the curses of dimensionality and history to be

lifted, which makes solving I-POMDP Lite very efficient, as explained below. Compared to existing game-theoretic frameworks [Hu and Wellman, 1998; Littman, 1994] which make strong assumptions of the other agent's behavior, our assumption is clearly less restrictive. Unlike the approximate I-POMDP techniques, it does not cause I-POMDP Lite to be subject to coarse approximation when solving larger problems, which can potentially result in better planning performance. Furthermore, by modeling and predicting the other agent's intention using nested MDP, I-POMDP Lite tends to overestimate its true sensing capability and can therefore achieve a more robust performance than I-PBVI using significantly less planning time under different partially observable environments (Section 5).

I-POMDP Lite Formulation. Our I-POMDP Lite framework constitutes an integration of the nested MDP for predicting the other agent's mixed strategy into a POMDP for tracking our agent's belief in partially observable environments. Naively, this can be achieved by extending the belief space to $\Delta(S \times V)$ (i.e., each belief b is now a probability distribution over the state-action space $S \times V$) and solving the resulting augmented POMDP. The size of representing each belief therefore becomes $\mathcal{O}(|S||V|)$ (instead of $\mathcal{O}(|S|)$), which consequently increases the cost of processing each belief (i.e., belief update). Fortunately, our I-POMDP Lite framework can alleviate this extra cost: By factorizing $b(s,v) = b(s)\ \widehat{\pi}^k_{\text{-}t}(s,v)$, the belief space over $S \times V$ can be reduced to one over S because the predictive probabilities $\widehat{\pi}^k_{\text{-}t}(s,v)$ (2) (i.e., predicted mixed strategy of the other agent) are derived separately in advance by solving nested MDPs. This consequently alleviates the curse of dimensionality pertaining to the use of interactive beliefs, as discussed in Section 1. Furthermore, such a reduction of the belief space decreases the time and space complexities and typically allows an optimal policy to be derived faster in practice: the space required to store n sampled beliefs is only $\mathcal{O}(n|S| + |S||V|)$ instead of $\mathcal{O}(n|S||V|)$.

Formally, I-POMDP Lite (for our agent t) is defined as a tuple $(S, U, V, O, T, Z, R, \widehat{\pi}^k_{\text{-}t}, \phi, b_0)$ where S is a set of all possible states of the environment; U and V are sets of all actions available to our agent t and the other agent -t, respectively; O is a set of all possible observations of our agent t; $T : S \times U \times V \times S \to [0,1]$ is a transition function that depends on the agents' joint actions; $Z : S \times U \times O \to [0,1]$ denotes the probability $Pr(o|s',u)$ of making observation $o \in O$ in state $s' \in S$ using our agent t's action $u \in U$; $R : S \times U \times V \to \mathbb{R}$ is the reward function of agent t; $\widehat{\pi}^k_{\text{-}t} : S \times V \to [0,1]$ denotes the predictive probability $Pr(v|s)$ of selecting action v in state s for the other agent -t and is derived using (2) by solving its nested MDPs at levels $0, \ldots, k-1$; $\phi \in (0,1)$ is a discount factor; and $b_0 \in \Delta(S)$ is a prior belief over the states of environment.

I-POMDP Lite Planning. Similar to solving POMDP (except for a few modifications), the optimal value function of I-POMDP Lite for our agent t satisfies the below Bellman equation:

$$V_{n+1}(b) = \max_u \Big(R(b,u) + \phi \sum_{v,o} Pr(v,o|b,u)\ V_n(b') \Big) \quad (5)$$

where our agent t's expected immediate payoff is

$$R(b,u) = \sum_{s,v} R(s,u,v)\ Pr(v|s)\ b(s) \quad (6)$$

and the belief update is given as

$$b'(s') = \beta\ Z(s',u,o) \sum_s T(s,u,v,s')\ Pr(v|s)\ b(s)\ .$$

Note that (6) yields an intuitive interpretation: The uncertainty over the state of the environment can be factored out of the prediction of the other agent -t's strategy by assuming that agent -t can fully observe the environment. Consequently, solving I-POMDP Lite (5) involves choosing the policy that maximizes the expected total reward with respect to the prediction of agent -t's mixed strategy using nested MDP. Like POMDP, the optimal value function $V_n(b)$ of I-POMDP Lite can be approximated arbitrarily closely (for infinite horizon) by a piecewise-linear and convex function that takes the form of a set V_n[3] of α vectors:

$$V_n(b) = \max_{\alpha \in V_n} (\alpha \cdot b)\ . \quad (7)$$

Solving I-POMDP Lite therefore involves computing the corresponding set of α vectors that can be achieved inductively: given a finite set V_n of α vectors, we can plug (7) into (5) to derive V_{n+1} (see Theorem 3 in Section 4). Similar to POMDP, the number of α vectors grows exponentially with the time horizon: $|V_{n+1}| = |U||V_n|^{|V||O|}$. To avoid this exponential blow-up, I-POMDP Lite inherits essential properties from POMDP (Section 4) that make it amenable to be solved by existing sampling-based algorithm such as PBVI [Pineau *et al.*, 2003] used here. The idea is to sample a finite set B of reachable beliefs (from b_0) to approximately represent the belief simplex, thus avoiding the need to generate the full belief reachability tree to compute the optimal policy. This alleviates the curse of history pertaining to the use of interactive beliefs (Section 1). Then, it suffices to maintain a single α vector for each belief point $b \in B$ that maximizes $V_n(b)$. Consequently, each backup step can be performed in polynomial time: $O(|U||V||O||B|^2|S|)$, as sketched below:

BACKUP(V_n, B)
1. $\Gamma^{u,*} \leftarrow \alpha^{u,*}(s) = \sum_v R(s,u,v)\ Pr(v|s)$
2. $\Gamma^{u,v,o} \leftarrow \forall \alpha'_i \in V_n\ \alpha^{u,v,o}_i(s) =$
$\qquad \phi\, Pr(v|s) \sum_{s'} Z(s',u,o)\, T(s,u,v,s')\ \alpha'_i(s')$
3. $\Gamma^u_b \leftarrow \Gamma^{u,*} + \sum_{v,o} \arg\max_{\alpha \in \Gamma^{u,v,o}} (\alpha \cdot b)$
4. Return $V_{n+1} \leftarrow \forall b \in B\ \arg\max_{\Gamma^u_b, \forall u \in U} (\Gamma^u_b \cdot b)$

Time Complexity. Given the set B of sampled beliefs, the cost of solving I-POMDP Lite is divided into two parts: (a) The cost of predicting the mixed strategy of the other agent using nested MDP (2) is $\mathcal{O}\big(kh|S|^2\big)$ (Section 2); (b) To determine the cost of approximately solving I-POMDP Lite with respect to this predicted mixed strategy, since each backup step incurs $\mathcal{O}\big(|U||V||O||B|^2|S|\big)$ time, solving I-POMDP Lite for h steps incurs $\mathcal{O}\big(h|U||V||O||B|^2|S|\big)$ time.

[3] With slight abuse of notation, the value function is also used to denote the set of corresponding α vectors.

By considering $|U|$, $|V|$, and $|O|$ as constants, the cost of solving I-POMDP Lite can be simplified to $\mathcal{O}\big(h|S||B|^2\big)$. Thus, the time complexity of solving I-POMDP Lite is $\mathcal{O}\big(h|S|(k|S|+|B|^2)\big)$, which is much less computationally demanding than the exponential cost of I-PF and I-PBVI (Section 1).

4 Theoretical Analysis

In this section, we prove that I-POMDP Lite inherits convergence, piecewise-linear, and convex properties of POMDP that make it amenable to be solved by existing sampling-based algorithms. More importantly, we show that the performance loss incurred by I-POMDP Lite is linearly bounded by the error of prediction of the other agent's strategy. This result also holds for that of nested MDP policy because I-POMDP Lite reduces to nested MDP under full observability.

Theorem 2 (Convergence) *Let V_∞ be the value function of I-POMDP Lite for infinite time horizon. Then, it is contracting/converging: $\|V_\infty - V_{n+1}\|_\infty \leq \phi\|V_\infty - V_n\|_\infty$.*

Theorem 3 (Piecewise Linearity and Convexity) *The optimal value function V_n can be represented as a finite set of α vectors: $V_n(b) = \max_{\alpha\in V_n}(\alpha \cdot b)$.*

We can prove by induction that the number of α vectors grows exponentially with the length of planning horizon; this explains why deriving the exact I-POMDP Lite policy is intractable in practice.

Definition 1 *Let π^*_{-t} be the true strategy of the other agent -t such that $\pi^*_{-t}(s,v)$ denotes the true probability $Pr^*(v|s)$ of selecting action $v \in V$ in state $s \in S$ for agent -t. Then, the prediction error is $\epsilon_p \triangleq \max_{v,s}|Pr^*(v|s) - Pr(v|s)|$.*

Definition 2 *Let $R_{\max} \triangleq \max_{s,u,v} R(s,u,v)$ be the maximum value of our agent t's payoffs.*

Theorem 4 (Policy Loss) *The performance loss δ_n incurred by executing I-POMDP Lite policy, induced w.r.t the predicted strategy $\widehat{\pi}^k_{-t}$ of the other agent -t using nested MDP (as compared to its true strategy π^*_{-t}), after n backup steps is linearly bounded by the prediction error ϵ_p :*

$$\delta_n \leq 2\epsilon_p|V|R_{\max}\left[\phi^{n-1} + \frac{1}{1-\phi}\left(1 + \frac{3\phi|O|}{1-\phi}\right)\right].$$

The above result implies that, by increasing the accuracy of the prediction of the other agent's strategy, the performance of the I-POMDP Lite policy can be proportionally improved. This gives a very strong motivation to seek better and more reliable techniques, other than our proposed nested MDP framework, for intention prediction. The formal proofs of the above theorems are provided in Appendix D of Hoang and Low [2013b].

5 Experiments and Discussion

This section first evaluates the empirical performance of nested MDP in a practical multi-agent task called *Intersection Navigation for Autonomous Vehicles* (INAV) (Section 5.1), which involves a traffic scenario with multiple cars coming from different directions (North, East, South, West) into an intersection and safely crossing it with minimum delay. Our goal is to implement an intelligent autonomous vehicle (**AV**) that cooperates well with human-driven vehicles (**HV**) to quickly and safely clear an intersection, in the absence of communication. Then, the performance of I-POMDP Lite is evaluated empirically in a series of partially observable stochastic games (POSGs) (Section 5.2). All experiments are run on a Linux server with two 2.2GHz Quad-Core Xeon E5520 processors and 24GB RAM.

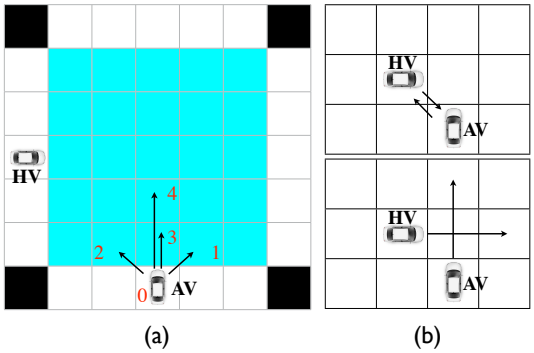

Figure 1: Intersection Navigation: (a) The road intersection modeled as a 7×7 grid (the black areas are not passable); and (b) accidents caused by cars crossing trajectories.

5.1 Performance of Nested MDP

In this task, the road intersection is modeled as a 7×7 grid, as shown in Fig. 1a. The autonomous car (**AV**) starts at the bottom row of the grid and travels North while a human-driven car (**HV**) starts at the leftmost column and travels to the East. Each car has five actions: Slow down (0), forward right (1), forward left (2), forward (3) and fast forward (4). Furthermore, it is assumed that 'slow down' has speed level 0, 'forward left', 'forward', and 'forward right' have speed level 1 while 'fast forward' has speed level 2. The difference in speed levels of two consecutive actions should be at most 1. In general, the car is penalized by the delay cost $D > 0$ for each executed action. But, if the joint actions of both cars lead to an accident by crossing trajectories or entering the same cell (Fig. 1b), they are penalized by the accident cost $C > 0$. The goal is to help the autonomous car to safely clear the intersection as fast as possible. So, a smaller value of D/C is desired as it implies a more rational behavior in our agent.

The above scenario is modeled using nested MDP, which requires more than 18000 states. Each state comprises the cells occupied by the cars and their current speed levels. The discount factor ϕ is set to 0.99. The delay and accident costs are hard-coded as $D = 1$ and $C = 100$. Nested MDP is used to predict the mixed strategy of the human driver and our car's optimal policy is computed with respect to this predicted mixed strategy. For evaluation, our car is run through 800 intersection episodes. The human-driven car is scripted with the following rational behavior: the human-driven car probabilistically estimates how likely a particular action will lead to an accident in the next time step, assuming that our car selects actions uniformly. It then forms a distribution over all actions such that most of the probability mass concentrates

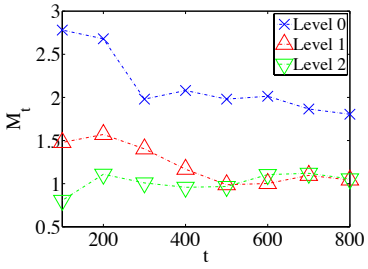

Figure 2: Performance comparison between nested MDPs at reasoning levels 0, 1, and 2.

on actions that least likely lead to an accident. Its next action is selected by sampling from this distribution.

We compare the performance of nested MDPs at reasoning levels $k = 0, 1, 2$. When $k = 0$, it is equivalent to the traditional MDP policy that treats the other car as environmental noise. During execution, we maintain a running average T_t (over the first t episodes) of the number of actions taken to clear an intersection and the number I_t of intersections experiencing accidents. The average ratio of the empirical delay is defined as $R^d_t = (T_t - T_{\min})/T_{\min} = T_t/T_{\min} - 1$ with $T_{\min} = 3$ (i.e., minimum delay required to clear the intersection). The empirical accident rate is defined as $R^c_t = I_t/t$. The average incurred cost is therefore $M_t = CR^c_t + DR^d_t$. A smaller M_t implies better policy performance.

Fig. 2 shows the results of the performance of the evaluated policies. It can be observed that the M_t curves of nested MDPs at reasoning levels 1 and 2 lie below that of MDP policy (i.e., reasoning level 0). So, nested MDP outperforms MDP. This is expected since our rationality assumption holds: Nested MDP's prediction is closer to the human driver's true intention and is thus more informative than the uniformly-distributed human driver's strategy assumed by MDP. Thus, we conclude that nested MDP is effective when the other agent's behavior conforms to our definition of rationality.

5.2 Performance of I-POMDP Lite

Specifically, we compare the performance of I-POMDP Lite vs. I-POMDP (at reasoning level $k = 1$) players under adversarial environments modeled as zero-sum POSGs. These players are tasked to compete against nested MDP and I-POMDP opponents at reasoning level $k = 0$ (i.e., respectively, MDP and POMDP opponents) whose strategies exactly fit the structural assumptions of I-POMDP Lite (Section 3) and I-POMDP (at $k = 1$), respectively. The I-POMDP player is implemented using I-PBVI, which is reported to be the best approximate I-POMDP technique [Doshi and Perez, 2008]. Each competition consists of 40 stages; the reward/penalty is discounted by 0.95 after each stage. The performance of each player, against its opponent, is measured by averaging its total rewards over 1000 competitions. Our test environment is larger than the benchmark problems in [Doshi and Perez, 2008]: There are 10 states, 3 actions, and 8 observations for each player. In particular, we let each of the first 6 states be associated with a unique observation with high probability. For the remaining 4 states, every disjoint pair of states is associated with a unique observation with high probability. Hence, the sensing capabilities of I-POMDP Lite and I-POMDP players are significantly weaker than that of the MDP opponent with full observability.

Table 1: I-POMDP's and I-POMDP Lite's performance against POMDP and MDP opponents with varying horizon lengths h ($|S| = 10, |A| = 3, |O| = 8$). '$*$' denotes that the program ran out of memory after 10 hours.

	POMDP	MDP	Time (s)	$\lvert IS' \rvert$
I-POMDP (h = 2)	13.33 ± 1.75	-37.88 ± 1.74	177.35	66110
I-POMDP (h = 3)	$*$	$*$	$*$	1587010
I-POMDP Lite (h = 1)	15.22 ± 1.81	15.18 ± 1.41	0.02	N.A.
I-POMDP Lite (h = 3)	17.40 ± 1.71	24.23 ± 1.54	0.45	N.A.
I-POMDP Lite (h = 8)	17.42 ± 1.70	24.66 ± 1.54	17.11	N.A.
I-POMDP Lite (h = 10)	17.43 ± 1.70	24.67 ± 1.55	24.38	N.A.

Table 1 shows the results of I-POMDP and I-POMDP Lite players' performance with varying horizon lengths. The observations are as follows: (a) Against a POMDP opponent whose strategy completely favors I-POMDP, both players win by a fair margin and I-POMDP Lite outperforms I-POMDP; (b) against a MDP opponent, I-POMDP suffers a huge loss (i.e., -37.88) as its structural assumption of a POMDP opponent is violated, while I-POMDP Lite wins significantly (i.e., 24.67); and (c) the planning times of I-POMDP Lite and I-POMDP appear to, respectively, grow linearly and exponentially in the horizon length.

I-POMDP's exponential blow-up in planning time is expected because its bounded interactive state space IS' increases exponentially in the horizon length (i.e., curse of dimensionality), as shown in Table 1. Such a scalability issue is especially critical to large-scale problems. To demonstrate this, Fig. 3b shows the planning time of I-POMDP Lite growing linearly in the horizon length for a large zero-sum POSG with 100 states, 3 actions, and 20 observations for each player; it takes about 6 and $1/2$ hours to plan for 100-step look-ahead. In contrast, I-POMDP fails to even compute its 2-step look-ahead policy within 12 hours.

It may seem surprising that I-POMDP Lite outperforms I-POMDP even when tested against a POMDP opponent whose strategy completely favors I-POMDP. This can be explained by the following reasons: (a) I-POMDP's exponential blow-up in planning time forbids it from planning beyond 3 look-ahead steps, thus degrading its planning performance; (b) as shown in Section 3, the cost of solving I-POMDP Lite is only polynomial in the horizon length and reasoning

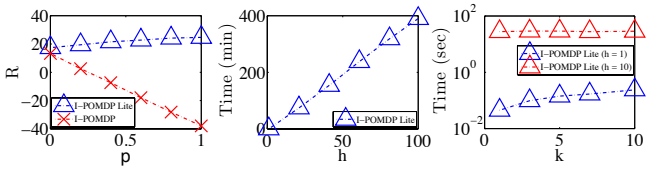

Figure 3: Graphs of (a) performance R of I-POMDP Lite and I-POMDP players against hybrid opponents ($|S| = 10, |A| = 3, |O| = 8$); (b) I-POMDP Lite's planning time vs. horizon length h in a large POSG ($|S| = 100, |A| = 3, |O| = 20$); and (c) I-POMDP Lite's planning time vs. reasoning level k for $h = 1$ and 10 ($|S| = 10, |A| = 3, |O| = 10$).

Table 2: I-POMDP's and I-POMDP Lite's performance against POMDP and MDP opponents with varying horizon lengths h ($|S| = 10, |A| = 3, |O| = 10$). '$*$' denotes that the program ran out of memory after 10 hours.

	POMDP	MDP	Time (s)	$\|IS'\|$
I-POMDP (h = 2)	5.70 ± 1.67	-9.62 ± 1.50	815.28	102410
I-POMDP (h = 3)	$*$	$*$	$*$	3073110
I-POMDP Lite (h = 1)	11.18 ± 1.75	20.25 ± 1.53	0.03	N.A.
I-POMDP Lite (h = 3)	14.89 ± 1.79	27.49 ± 1.53	0.95	N.A.
I-POMDP Lite (h = 8)	14.99 ± 1.79	26.91 ± 1.55	24.10	N.A.
I-POMDP Lite (h = 10)	15.01 ± 1.79	26.91 ± 1.55	33.74	N.A.

depth, thus allowing our player to plan with a much longer look-ahead (Fig. 3b) and achieve substantially better planning performance; and (c) with reasonably accurate observation and transition models, Theorem 1 indicates that the strategy of the MDP opponent (i.e., nested MDP at reasoning level 0) is likely to approximate that of the true POMDP opponent closely, thus reducing the degree of violation of I-POMDP Lite's structural assumption of a nested MDP opponent. Such an assumption also seems to make our I-POMDP Lite player overestimate the sensing capability of an unforeseen POMDP opponent and consequently achieve a robust performance against it. On the other hand, the poor performance of I-POMDP against a MDP opponent is expected because I-POMDP's structural assumption of a POMDP opponent is likely to cause its player to underestimate an unforeseen opponent with superior sensing capability (e.g., MDP) and therefore perform badly against it. In contrast, I-POMDP Lite performs significantly better due to its structural assumption of a nested MDP opponent at level 0, which matches the true MDP opponent exactly.

Interestingly, it can be empirically shown that when both players' observations are made more informative than those used in the previous experiment, the performance advantage of I-POMDP Lite over I-POMDP, when tested against a POMDP opponent, increases. To demonstrate this, we modify the previous zero-sum POSG to involve 10 observations (instead of 8) such that every state (instead of a disjoint pair of states) is associated with a unique observation with high probability (i.e., ≥ 0.8); the rest of the probability mass is then uniformly distributed among the other observations. Hence, the sensing capabilities of I-POMDP Lite and I-POMDP players in this experiment are much better than those used in the previous experiment and hence closer to that of the MDP opponent with full observability. Table 2 summarizes the results of I-POMDP Lite's and I-POMDP's performance when tested against the POMDP and MDP opponents in the environment described above.

To further understand how the I-POMDP Lite and I-POMDP players perform when the sensing capability of an unforeseen opponent varies, we set up another adversarial scenario in which both players pit against a hybrid opponent: At each stage, with probability p, the opponent knows the exact state of the game (i.e., its belief is set to be peaked at this known state) and then follows the MDP policy; otherwise, it follows the POMDP policy. So, a higher value of p implies better sensing capability of the opponent. The environment settings are the same as those used in the first experiment, that is, 10 states, 8 observations and 3 actions for each player (Table 1). Fig. 3a shows the results of how the performance, denoted R, of I-POMDP Lite and I-POMDP players vary with p: I-POMDP's performance decreases rapidly as p increases (i.e., opponent's strategy violates I-POMDP's structural assumption more), thus increasing the performance advantage of I-POMDP Lite over I-POMDP. This demonstrates I-POMDP Lite's robust performance when tested against unforeseen opponents whose sensing capabilities violate its structural assumption.

To summarize the above observations, (a) in different partially observable environments where the agents have reasonably accurate observation and transition models, I-POMDP Lite significantly outperforms I-POMDP (Tables 1, 2 and Fig. 3a); and (b) interestingly, it can be observed from Fig. 3a that when the sensing capability of the unforeseen opponent improves, the performance advantage of I-POMDP Lite over I-POMDP increases. These results consistently demonstrate I-POMDP Lite's robust performance against unforeseen opponents with varying sensing capabilities. In contrast, I-POMDP only performs well against opponents whose strategies completely favor it, but its performance is not as good as that of I-POMDP Lite due to its limited horizon length caused by the extensive computational cost of modeling the opponent. Unlike I-POMDP's exponential blow-up in horizon length h and reasoning depth k (Section 1), I-POMDP Lite's processing cost grows linearly in both h (Fig. 3b) and k (Fig. 3c). When $h = 10$, it can be observed from Fig. 3c that I-POMDP Lite's overall processing cost does not change significantly with increasing k because the cost $\mathcal{O}\big(kh|S|^2\big)$ of predicting the other agent's strategy with respect to k is dominated by the cost $\mathcal{O}\big(h|S||B|^2\big)$ of solving I-POMDP Lite for large h (Section 3).

6 Conclusion

This paper proposes the novel nested MDP and I-POMDP Lite frameworks [Hoang and Low, 2012], which incorporate the cognitive hierarchy model of games [Camerer *et al.*, 2004] for intention prediction into the normative decision-theoretic POMDP paradigm to address some practical limitations of existing planning frameworks for self-interested MAS such as computational impracticality [Doshi and Perez, 2008] and restrictive equilibrium theory of agents' behavior [Hu and Wellman, 1998]. We have theoretically guaranteed that the performance losses incurred by our I-POMDP Lite policies are linearly bounded by the error of intention prediction. We have empirically demonstrated that I-POMDP Lite performs significantly better than the state-of-the-art planning algorithms in partially observable stochastic games. Unlike I-POMDP, I-POMDP Lite's performance is very robust against unforeseen opponents whose sensing capabilities violate the structural assumption (i.e., of a nested MDP opponent) that it has exploited to achieve significant computational gain. In terms of computational efficiency and robustness in planning performance, I-POMDP Lite is thus more practical for use in larger-scale problems.

Acknowledgments. This work was supported by Singapore-MIT Alliance Research & Technology Subaward Agreements No. 28 R-252-000-502-592 & No. 33 R-252-000-509-592.

References

[Camerer *et al.*, 2004] C. F. Camerer, T. H. Ho, and J. K. Chong. A cognitive hierarchy model of games. *Quarterly J. Economics*, 119(3):861–898, 2004.

[Chalkiadakis and Boutilier, 2003] Georgios Chalkiadakis and Craig Boutilier. Coordination in multiagent reinforcement learning: A Bayesian approach. In *Proc. AAMAS*, pages 709–716, 2003.

[Doshi and Gmytrasiewicz, 2009] P. Doshi and P. Gmytrasiewicz. Monte Carlo sampling methods for approximating interactive POMDPs. *JAIR*, pages 297–337, 2009.

[Doshi and Perez, 2008] P. Doshi and D. Perez. Generalized point based value iteration for interactive POMDPs. In *Proc. AAAI*, pages 63–68, 2008.

[Gigerenzer and Selten, 2002] G. Gigerenzer and R. Selten. *Bounded Rationality*. MIT Press, 2002.

[Gmytrasiewicz and Doshi, 2005] P. J. Gmytrasiewicz and P. Doshi. A framework for sequential planning in multiagent settings. *JAIR*, 24:49–79, 2005.

[Hansen *et al.*, 2004] E. A. Hansen, D. S. Bernstein, and S. Zilberstein. Dynamic programming for partially observable stochastic games. In *Proc. AAAI*, pages 709–715, 2004.

[Hoang and Low, 2012] T. N. Hoang and K. H. Low. Intention-aware planning under uncertainty for interacting with self-interested, boundedly rational agents. In *Proc. AAMAS*, pages 1233–1234, 2012.

[Hoang and Low, 2013a] T. N. Hoang and K. H. Low. A general framework for interacting Bayes-optimally with self-interested agents using arbitrary parametric model and model prior. In *Proc. IJCAI*, 2013.

[Hoang and Low, 2013b] T. N. Hoang and K. H. Low. Interactive POMDP Lite: Towards practical planning to predict and exploit intentions for interacting with self-interested agents. arXiv:1304.5159, 2013.

[Hu and Wellman, 1998] J. Hu and M. P. Wellman. Multiagent reinforcement learning: Theoretical framework and an algorithm. In *Proc. ICML*, pages 242–250, 1998.

[Littman, 1994] M. L. Littman. Markov games as a framework for multi-agent reinforcement learning. In *Proc. ICML*, pages 157–163, 1994.

[Nair and Tambe, 2003] R. Nair and M. Tambe. Taming decentralized POMDPs: Towards efficient policy computation for multiagent settings. In *Proc. IJCAI*, pages 705–711, 2003.

[Ng *et al.*, 2010] B. Ng, C. Meyers, K. Boakye, and J. Nitao. Towards applying interactive POMDPs to real-world adversary modeling. In *Proc. IAAI*, pages 1814–1820, 2010.

[Pineau *et al.*, 2003] J. Pineau, G. Gordon, and S. Thrun. Point-based value iteration: An anytime algorithm for POMDPs. In *Proc. IJCAI*, pages 1025–1032, 2003.

[Rathnasabapathy *et al.*, 2006] B. Rathnasabapathy, P. Doshi, and P. Gmytrasiewicz. Exact solutions of interactive POMDPs using behavioral equivalence. In *Proc. AAMAS*, pages 1025–1032, 2006.

[Seuken and Zilberstein, 2007] S. Seuken and S. Zilberstein. Memory-bounded dynamic programming for DEC-POMDPs. In *Proc. IJCAI*, pages 2009–2015, 2007.

[Spaan *et al.*, 2011] M. T. J. Spaan, F. A. Oliehoek, and C. Amato. Scaling up optimal heuristic search in DEC-POMDPs via incremental expansion. In *Proc. IJCAI*, pages 2027–2032, 2011.