Individual Planning in Open and Typed Agent Systems

Muthukumaran Chandrasekaran
University of Georgia
Athens, GA 30602

Adam Eck
University of Nebraska
Lincoln, NE 68588

Prashant Doshi
University of Georgia
Athens, GA 30602

Leenkiat Soh
University of Nebraska
Lincoln, NE 68588

Abstract

Open agent systems are multiagent systems in which one or more agents may leave the system at any time possibly resuming after some interval and in which new agents may also join. Planning in such systems becomes challenging in the absence of inter-agent communication because agents must predict if others have left the system or new agents are now present to decide on possibly choosing a different line of action. In this paper, we prioritize open systems where agents of differing types may leave and possibly reenter but new agents do not join. With the help of a realistic domain – wildfire suppression – we motivate the need for individual planning in open environments and present a first approach for robust decision-theoretic planning in such multiagent systems. Evaluations in domain simulations clearly demonstrate the improved performance compared to previous methods that disregard the openness.

1 INTRODUCTION

The past year has been witness to one of the worst seasons of wildland fires, here onwards referred to as wildfires, on record in the United States. There were more than fifty thousand wildfires that burned more than nine million acres of wildland. Both ground and various types of aerial firefighting units are often deployed in suppressing these fires. Consider the decision-making task of a small ground unit of firefighters. As these fires are large, a unit needs to coordinate with others to focus their resources on the same area and make a difference. However, units may run out of suppressants (such as water and chemicals) or suffer from exhaustion causing them to temporarily leave. Consequently, wildfire fighting units form an *open* and *typed* multiagent system and a unit's decision making about its course of action becomes challenging if a leaving unit is unable to radio its intent to temporarily disengage.

Open agent systems such as the one described above are characterized by one or more interacting agents leaving the system at any time and possibly resuming after some interval, or new agents joining the system [1]. We refer to this characteristic as *agent openness*. A second form of openness is exhibited by a system when the types of agents alter at any time perhaps just briefly; we refer to this as *type openness*. The wildfire suppressing example presented above exhibits agent openness but not type openness.

In this paper, we prioritize systems exhibiting agent openness and further limit our attention to systems where agents may disengage at any time and possibly reenter the system but new agents do not enter the system. We are interested in how an individual agent, for example the ground firefighting unit, should plan its actions in such open and typed multiagent systems. A perspectivistic approach makes this investigation broadly applicable to cooperative, noncooperative and mixed settings all of which may exhibit agent openness.

Previous methods for individual decision-theoretic planning such as algorithms for the well-known interactive partially observable Markov decision process (I-POMDP) [2, 3] are well suited for typed systems but do not model agent openness so far. Similarly, algorithms for joint cooperative planning in frameworks such as the decentralized POMDP [4] are not easily amended for open agent systems. As such, there is a marked gap in the literature on principled planning for open systems. We present a first approach for modeling and planning in the context of agent openness when the physical state may not be perfectly observed. In keeping with our objective of individual planning and the presence of agents of various types, we generalize the I-POMDP-Lite framework [5] to allow for agent openness. This framework is more efficient than the general I-POMDP because it ascribes a nested MDP to model others rather than a belief hierarchy. We utilize a graph to model the interaction structure between various agents and extend the joint state to model the event that neighboring agents could have disengaged. In the absence of communication, we show how the

agent's unexpected observations allow it to correct its model of the other agent *post hoc* after the agent has left or has reentered. Alternately, a *proactive* approach that seeks to predict when agent may disengage or reenter should exhibit improved benefit. However, the subject agent may not know how factors relevant to others' decisions to leave or reenter evolve.

The generalized I-POMDP-Lite is utilized to model the problem domain of wildfire suppression exhibiting open and typed agent systems. We continue with Hoang and Low's [5] use of interactive point-based value iteration to scalably plan and extend it appropriately to the generalized I-POMDP-Lite. Evaluations in simulations of the domain clearly demonstrate not only the improved performance of the individual agent but also the performance of the entire open system at a macro level in comparison to planning that disregards agent openness.

2 RELATED WORK

Agent openness is a challenging property of realistic multiagent systems that has been identified at various points in the literature. Shehory [1] noted that the openness of an agent system refers to the ability of introducing additional agents into the system in excess to the agents that comprise it initially. Calmet et al. [6] also studied openness in societies of agents where an open society is one that is open to new agents either with no definite goal or with goals not exceedingly relevant to the society. Both definitions focused on the system and software architecture to support openness.

Recently, additional properties for agent openness have been reported. Jamroga et al. [7] defined the degree of openness of multi-agency as the complexity of the minimal transformation that the system must undergo in order to add a new agent to the system or remove an existing one from the system. Jumadinova et al. [8] and Chen et al. [9] extended the notion of openness to include both agent openness and task openness to model the dynamic nature of the agents and tasks in the environment. They considered fluctuations in the availability of agents needed to perform tasks, as well as dynamic changes in the type of tasks that appear over time. In both papers, the degree of openness is defined as the rate at which agents/tasks join and leave the environment.

Relevant to modeling individual agents in open environments, Huynh et al. [10] studied the problem of developing trust and reputation models in open agent systems to enable agents (owned by a variety of stakeholders) to assess the quality of their peers' likely performance. Similarly, Pinyol and Sabater-Mir [11] studied, for open environments where agents' intentions are unknown, how to control the interactions among the agents in order to protect good agents from fraudulent entities, or to help agents find trustworthy or reputable agents.

In this paper, we adopt the notion of dynamic agent openness defined by Shehory, extended in Jumadinova et al. and Chen et al. Similar to Huynh et al. and Pinyol and Sabater-Mir, we are also interested in developing a solution to enable an agent to model its transient neighbors in open environments. However, our problem and approach differ in that we are interested in modeling *how neighbors will behave over time* (i.e., predicting what actions they might take, as well as their future presence in the environment which directly impacts when and how they might work together with or against an agent), instead of determining how reliable a neighbor might be. Similar to Jumadinova et al. and Chen et al., we also seek to design agents capable of strategic, self-interested reasoning, but we do so from the decision-theoretic perspective grounded in the tradition of Markov decision problems with an added focus on modeling peer behavior in order to plan and perform actions as a best response to the expected behavior of peers.

Finally, we note that ad hoc cooperation – coming together of multiple agents on the fly to meet a goal [12] – is just one characteristic of an open agent system. As this paper's focus is instead on agents dynamically departing the system and reentering, we do not discuss the emerging literature on online planning for ad hoc teamwork.

3 BACKGROUND

I-POMDP-Lite mitigates the complexity of I-POMDPs by predicting other agent's actions using a nested MDP; this assumes that the other perfectly observes the physical state. A nested MDP [5] is a scalable framework for individual planning in multiagent systems where the physical state and others' models are perfectly observable to each agent. It is defined as a tuple for agent i:

$$\mathcal{M}_{i,l} \triangleq \langle S, A, T_i, R_i, \{\pi_{j,d}, \pi_{k,d}, \ldots, \pi_{z,d}\}_{d=0}^{l-1}, OC_i \rangle$$

where:

- S is the set of physical states of the interacting agent system. The space may be factored as, $S = X_1 \times X_2 \times \ldots \times X_k$, where $X_1, \ldots, X_k$ are $k > 0$ factors;
- $A = A_i \times A_j \times \ldots \times A_z$ is the set of joint actions of all interacting agents in the system;
- Transition of a state due to the joint actions to another state may be stochastic and the transition function is defined as, $T_i : S \times A \times S \rightarrow [0, 1]$. The transition probabilities may be conditionally factored based on the factorization of the state space;
- R_i is the reward function of agent i that depends on the state and joint actions, $R_i : S \times A \rightarrow \mathbb{R}$;
- $\{\pi_{j,d}, \pi_{k,d}, \ldots, \pi_{z,d}\}_{d=0}^{l-1}$ is the set of other agents j, k, …, z reasoning models at all levels from 0 to $l-1$. Each of these models is a policy which is a mapping from states

to distributions over actions and is obtained by solving a nested MDP for the agent at that level. However, a level-0 reasoning model is a uniform distribution over an agent's actions;

- OC_i is i's optimality criterion. In this paper, we utilize a finite horizon H with discount factor $\gamma \in (0, 1)$.

Analogous to MDPs, we may associate a horizon $0 < h \leq H$ value function with $\mathcal{M}_{i,l}$ that extends the standard Bellman equation. Let $A_{-i} = A_j \times A_k \times \ldots \times A_z$.

$$V^h_{i,l}(s) = \max_{a_i \in A_i} \sum_{\mathbf{a}_{-i} \in A_{-i}} \prod_{-i \in \{j,k,\ldots,z\}} \hat{\pi}_{-i,l-1}(s, a_{-i}) \times Q^h_{i,l}(s, a_i, \mathbf{a}_{-i}) \quad (1)$$

Here, $Q^h_{i,l}(s, a_i, \mathbf{a}_{-i})$ is defined recursively:

$$Q^h_{i,l}(s, a_i, \mathbf{a}_{-i}) = R_i(s, a_i, \mathbf{a}_{-i}) + \gamma \sum_{s' \in S} T_i(s, a_i, \mathbf{a}_{-i}, s') \times V^{h-1}_{i,l}(s') \quad (2)$$

Furthermore, $\hat{\pi}_{-i,l-1}$ in Eq. 1 is defined as a mixed strategy that has a distribution over reasoning models at all levels up to $l-1$. If $l-1=0$ in Eq. 1 then $\hat{\pi}_{-i,l-1}$ is a uniform distribution over the other agent's actions.

$$\hat{\pi}_{j,l}(s, a_j) \triangleq \begin{cases} \sum_{d=0}^{l} Pr(d)\ \pi_{j,d}(s, a_j) & l \geq 1 \\ \frac{1}{|A_j|} & l = 0 \end{cases} \quad (3)$$

Policy $\pi_{j,d}(s, a_j)$ is obtained by solving the nested MDP of agent j at level d, which involves optimizing the corresponding value function similar to Eq. 1. Let Opt_j be the set of j's actions that optimizes it. Then, $\pi_{j,d}(s, a_j) = \frac{1}{|Opt_j|}$ if $a_j \in |Opt_j|$ otherwise $\pi_{j,d}(s, a_j)$ is 0. Distribution $Pr(d)$ on the nesting depth up to l is typically uniform but may also be learned from data as well.

With all agents modeling each other, solution of a nested MDP proceeds bottom up. Level-0 models of all agents default to uniform distributions. These are utilized in solving level-1 nested MDPs, $\mathcal{M}_{i,1}, \mathcal{M}_{j,1}, \ldots, \mathcal{M}_{z,1}$. Both level-0 and -1 solutions are utilized in solving nested MDPs at level 2; and so on up to level l. Consequently, in an N-agent system we solve $N-1$ models of others at any level and a total of $(N-1)l$ models. This is *linear* in both the number of nesting levels l and the number of agents, and scales well with both. If all N agents plan using nested MDPs then a total of $\mathcal{O}(N^2 l)$ such models are solved.

For individual planning in situations where the physical state is not perfectly observable to the subject agent i although the reasoning models of other agents are known and are supposed to possess the capability to observe the state perfectly, Hoang and Low [5] present the I-POMDP-Lite framework.

$$\text{I-POMDP}^{\mathcal{L}}_{i,l} \triangleq \langle S, A, \Omega_i, T_i, O_i, R_i, \{\mathcal{M}_{j,l-1}, \mathcal{M}_{k,l-1}, \ldots, \mathcal{M}_{z,l-1}\}, OC_i \rangle$$

Parameters S, A, T_i and R_i are as defined previously in the nested MDP framework. Ω_i is the set of agent i's observations and O_i is the observation function, which models the level of noise in the observations: $O_i : S \times A_i \times \Omega_i \rightarrow [0, 1]$. Notice that the observation distribution is conditionally independent of other agents' actions. $\{\mathcal{M}_{j,l-1}, \mathcal{M}_{k,l-1}, \ldots, \mathcal{M}_{z,l-1}\}$ are the nested MDPs of various interacting agents, and OC_i is the optimality criterion that may include a discount factor and an initial belief, b^0_i, over the state space.

Analogous to POMDPs, an agent maintains a belief over the states and the planning method associates a value function with the belief:

$$V^h_{i,l}(b_i) = \max_{a_i \in A_i} \left(\rho_i(b_i, a_i) + \gamma \sum_{s' \in S, o_i \in \Omega_i} \mathcal{T}_i^{a_i, o_i}(s', o_i | b_i, a_i) \times V^{h-1}_{i,l}(b'_i) \right) \quad (4)$$

where,

$$\rho_i(b_i, a_i) = \sum_{s \in S} \sum_{\mathbf{a}_{-i} \in A_{-i}} \prod_{-i \in \{j,k,\ldots,z\}} \pi_{-i,l-1}(s, a_{-i}) \times R_i(s, a_i, \mathbf{a}_{-i})\ b_i(s)$$

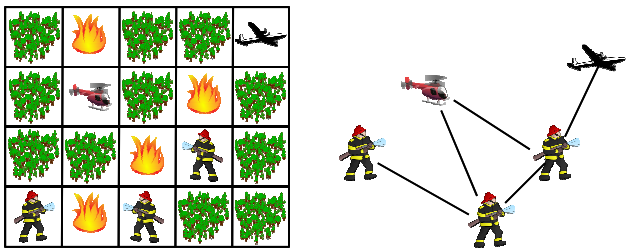
Policies $\pi_{-i,l-1}(s, a_{-i})$, $-i \in \{j, k, \ldots, z\}$ are solutions of the other agents' nested MDPs; and b'_i denotes the updated belief, $Pr(s'|o_i, a_i, \mathbf{a}_{-i}, b_i) \propto O_i(s', a_i, o_i) \times \sum_{s \in S} T_i(s, a_i, \mathbf{a}_{-i}, s')\ b_i(s)$.

Solution of I-POMDP$^{\mathcal{L}}_{i,l}$ requires solving the nested MDPs that are a part of its definition to obtain the policies. As we mentioned previously, this proceeds bottom up. At the top most level only, a POMDP is solved by decomposing the value function given in Eq. 4 into an inner product between a set of alpha vectors and the belief. While the total number of models that are solved remain linear in the nesting level and the number of agents, the computational complexity is higher because of the presence of a POMDP.

4 INDIVIDUAL PLANNING WITH AGENT OPENNESS

Planning that can assist fighting wildfires must deal with the event that units run out of suppressants – some types of units run out more quickly than others – due to which units temporarily leave the theater and thus the agent system. We seek to reason about the agent openness found in such environments as part of the individual planning in a principled way.

Systems of many agents in the real world often exhibit interaction structure. Specifically, not all agents interact with one another; rather, interactions often happen among small subgroups of agents.



(a) (b)

Figure 1: (a) An example wildfire scenario with 5 firefighting units of three types situated in a 4×5 grid of forestland. (b) Firefighting units must often coordinate on suppressing fires. This coordination overlays an interaction graph.

A well-known data structure that explicates the interaction structure is an *interaction graph*, which is an undirected graph whose nodes represent agents and the absence of an edge between two agents indicates that the reward of each of the two agents is not dependent on any action of the other agent. Interaction structure may be exploited during planning for computational scalability [13, 14, 15]. We motivate the interaction structure using an example:

Example 1. *Figure 1(a) illustrates an example wildfire suppressing scenario that consists of ground and two types of aerial firefighting units. Units must coordinate on adjacent or diagonally located fires to gradually suppress them and prevent them from spreading. This coordination overlays an interaction graph that is shown in Fig. 1(b). Notice that the graph is not a clique and thus exhibits structure.*

Each vertex in the graph denotes an agent i in the set $\mathcal{N}$ of agents and an edge between a pair of agents whose individual payoffs depend on each other's actions. Let $\nu(i)$ be the set of nodes that are directly linked by an edge to node i. We refer to $\nu(i)$ as the set of i's neighboring agents.

4.1 Post Hoc Reasoning

Let $\dot{\mathbf{X}}$ be the set of distinguished state factor(s) in S whose value determines whether other agent $j \in \nu(i)$ temporarily leaves the network. For example, this variable could reflect j's suppressant level. If i determines that j has left the network, i replaces j's predicted actions – using j's policy obtained by solving its nested MDP $\mathcal{M}_{j,l-1}$ – with a no op action from then onward during which the agent does not act. [1] Consequently, A_j is replaced with $A_j \cup \{\text{no op}\}$. This is beneficial because we need not change the definitions of agent i's transition, observation and reward functions when an agent leaves the network if no op is already in A_j. Otherwise, these are modified to model the implications of j's no operation to allow for agent openness.

[1] A caveat of this approach is that the agent's absence is observationally equivalent with the agent intentionally not acting.

However, the problem of predicting when an agent has left the network remains challenging because the state is partially observable – the amount of j's suppressant cannot be directly observed as we preclude communication between the separated units. Similarly, the problem of predicting if and when an agent has resumed its activities is also challenging.

Define joint probability $\mathcal{T}_i^{a_i,o_i}(s', o_i|a_i, b_i)$ as,

$$\mathcal{T}_i^{a_i,o_i}(s', o_i|a_i, b_i) = \sum_{s \in S} b_i(s) \sum_{\mathbf{a}_{-i} \in A_{-i}} T_i(s, a_i, \mathbf{a}_{-i}, s') \\ O_i(s', a_i, \mathbf{a}_{-i}, o_i) \prod_{-i \in \{j,k,\ldots,z\}} \pi_{-i,l-1}(s, a_{-i})$$

We make the following key observation that facilitates progress in this challenging task:

Observation 1. Post hoc $\mathcal{T}_i^{a_i,o_i}(s', o_i|a_i, b_i)$ *immediately after a neighboring agent has left the system will be small but will generally increase with time until the agent reenters.*

While agent j may abruptly leave the system at time step t, the planning agent continues to predict j's actions as if it were part of the system until the observation at time $t+1$ reveals that the state did not transition as expected. In other words, joint $\mathcal{T}_i^{a_i,o_i}(s', o_i|a_i, b_i)$ will be small because next state s' that is obtained by predicting j's action incorrectly will have low likelihood given observation o_i.

Nevertheless, observation o_i when used in the belief update to obtain b_i' will cause the probability mass in b_i to shift to states that make o_i more likely. These are likely to be states at which $\pi_{j,l-1}(s', \text{no op})$ is high; i.e., j is not performing any significant action because j has left the system. With more such observations that support the fact that j has left the system, more probability mass in the updated beliefs settles on states at which j is predicted to not perform any significant action. Therefore, joint $\mathcal{T}_i^{a_i,o_i}(s', o_i|a_i, b_i)$ will start rising until another such event occurs. We illustrate this observation:

Example 2. *Let firefighting unit j exit a team that consists of units i and j who are coordinating on suppressing a high intensity wildfire. Unit i expects the intensity of the fire to continue reducing in the next time step but instead observes that the intensity remained the same as before. This low probability observation makes i subsequently believe that perhaps j is not fighting the fire anymore (because it may have left the system); a belief that gets strengthened further as the fire continues to burn at the same intensity despite i fighting it. When its predictions of j performing* no ops *are sufficiently certain, i may choose to coordinate on a different wildfire with another unit.*

Observation 1 continues to hold when $\nu(i)$ has two or more agents but i may not be able to pinpoint which agent has exited.

Moving forward, let agent j reenter the system and resume its actions. Again, agent i may experience a phenomenon similar to that described in Observation 1 where *post hoc* $\mathcal{T}_i(s', o_i|a_i, b_i)$ drops because the observations do not support the next predicted state. This is because i is attributing no op to j despite j having reentered. However, persistent observations will shift the probability mass in i's belief to states at which j is predicted to be performing actions other than no op thereby modeling the fact that j is active again.

To illustrate, if unit i continued fighting the same wildfire as before, it may suddenly witness the intensity reducing significantly. This is indicative of the fact that unit j is active again and i's revised beliefs will emphasize those states at which j could also be suppressing the fire. On reentering, an agent can be expected to remain committed to fighting the fire if the intensity is steadily reducing, until the fire is suppressed or the agent's suppressant becomes low.

The observation below summarizes this subsection:

Observation 2. *Decision-theoretic planning that integrates modeling behaviors of other agents and a Bayesian belief update can reason about agent openness post hoc and plan accordingly with minimal extension.*

However, the limitation is that the adaptation of planning to the dynamic openness is delayed due to the *post hoc* reasoning.

4.2 Predicting Agent Openness

A better way to act in open agent systems would be to predict when a neighboring agent leaves or reenters the system. Presuming that the agent's departure is a policy-guided behavior and the agent's policy is known, we must predict changes in the distinguished state factors $\dot{\mathbf{X}}$ that may cause the other agent to leave the system. However, the main challenge is that the subject agent is usually uninformed about how these factors evolve over time.

For example, the rate at which the other firefighting unit consumes its suppressant is typically not known and may not be observed due to the separation between the two units. Nevertheless, observations related to the unit leaving the system over time provide information from which the rate could be gradually learned and utilized in the prediction.

Let $\dot{T}_i(\dot{\mathbf{x}}, a_i, \mathbf{a}_{-i}, \dot{\mathbf{x}}')$ be the transition distribution for $\dot{\mathbf{X}}$ given i's action a_i, neighbors' joint actions $\mathbf{a}_{-i}$, and previous value of the factor $\dot{\mathbf{x}} \in \dot{\mathbf{X}}$. For notational convenience, we assume that $\dot{T}_i$ may be factored out from function T_i. Subsequently, predicting when neighboring agents $j \in \nu(i)$ are likely to leave the system is dependent on knowing $\dot{T}_i(\dot{\mathbf{x}}, a_i, \mathbf{a}_{-i}, \dot{\mathbf{x}}')$ for all pairs of state factors and joint actions.

Our approach is Bayesian; it involves explicitly modeling the uncertainty over the distribution, updating it over time based on expected next states and utilizing it in the offline planning.

We may model the uncertainty over the distribution $\dot{T}_i(\dot{\mathbf{x}}, a_i, \mathbf{a}_{-i}, \cdot)$ as a Dirichlet process (DP), and the uncertainty over all such distributions as a system of Dirichlet processes. Formally, $\dot{T}_i(\dot{\mathbf{x}}, a_i, \mathbf{a}_{-i}, \cdot) \sim DP(n, C)$, where n is a positive integer and C is a distribution over $\dot{\mathbf{X}}$. Let factor(s) $\dot{\mathbf{X}}$ assume values $\{\dot{\mathbf{x}}_1, \dot{\mathbf{x}}_2, \ldots, \dot{\mathbf{x}}_{|\dot{\mathbf{X}}|}\}$, then

$$\Big(\dot{T}_i(\dot{\mathbf{x}}, a_i, \mathbf{a}_{-i}, \dot{\mathbf{x}}_1), \dot{T}_i(\dot{\mathbf{x}}, a_i, \mathbf{a}_{-i}, \dot{\mathbf{x}}_2), \ldots, \dot{T}_i(\dot{\mathbf{x}}, a_i, \mathbf{a}_{-i}, \dot{\mathbf{x}}_{|\dot{\mathbf{X}}|})\Big) \sim Dir(n \cdot \frac{c^{\dot{\mathbf{x}}\mathbf{a}\dot{\mathbf{x}}_1}}{n}, n \cdot \frac{c^{\dot{\mathbf{x}}\mathbf{a}\dot{\mathbf{x}}_2}}{n}, \ldots, n \cdot \frac{c^{\dot{\mathbf{x}}\mathbf{a}\dot{\mathbf{x}}_{|\dot{\mathbf{X}}|}}}{n}) \tag{5}$$

where $c^{\dot{\mathbf{x}}\mathbf{a}\dot{\mathbf{x}}_1}$ is the number of samples where transition $(\dot{\mathbf{x}}, a_i, \mathbf{a}_{-i}, \dot{\mathbf{x}}_1)$ occurs, and analogously for others; $n \triangleq \sum_{q=1}^{|\dot{\mathbf{X}}|} c^{\dot{\mathbf{x}}\mathbf{a}\dot{\mathbf{x}}_q}$ is the total number of samples. A Dirichlet process has the appealing property that the mean of its marginal, $E[\dot{T}_i(\dot{\mathbf{x}}, a_i, \mathbf{a}_{-i}, \dot{\mathbf{x}}_1)] = \frac{c^{\dot{\mathbf{x}}\mathbf{a}\dot{\mathbf{x}}_1}}{n}$ and the concentration parameter n inversely impacts the variance.

Let us obtain a sequence of n' next states $\{\dot{\mathbf{x}}'_1, \ldots, \dot{\mathbf{x}}'_{n'}\}$ given the current state $\dot{\mathbf{x}}$ and actions $a_i, \mathbf{a}_{-i}$ in independent draws. Then the posterior distributions become,

$$\Big(\dot{T}_i(\dot{\mathbf{x}}, a_i, \mathbf{a}_{-i}, \dot{\mathbf{x}}_1), \dot{T}_i(\dot{\mathbf{x}}, a_i, \mathbf{a}_{-i}, \dot{\mathbf{x}}_2), \ldots, \dot{T}_i(\dot{\mathbf{x}}, a_i, \mathbf{a}_{-i}, \dot{\mathbf{x}}_{|\dot{\mathbf{X}}|})\Big) \,|\dot{\mathbf{x}}'_1, \ldots, \dot{\mathbf{x}}'_{n'} \sim Dir\,(n + n' \cdot \frac{c^{\dot{\mathbf{x}}\mathbf{a}\dot{\mathbf{x}}_1} + \sum_{q=1}^{n'} \delta_{\dot{\mathbf{x}}_1}(\dot{\mathbf{x}}_q)}{n + n'}, n + n' \cdot \frac{c^{\dot{\mathbf{x}}\mathbf{a}\dot{\mathbf{x}}_2} + \sum_{q=1}^{n'} \delta_{\dot{\mathbf{x}}_2}(\dot{\mathbf{x}}_q)}{n + n'}, \ldots, n + n' \cdot \frac{c^{\dot{\mathbf{x}}\mathbf{a}\dot{\mathbf{x}}_{|\dot{\mathbf{X}}|}} + \sum_{q=1}^{n'} \delta_{\dot{\mathbf{x}}_{|\dot{\mathbf{X}}|}}(\dot{\mathbf{x}}_q)}{n + n'}\Big) \tag{6}$$

where $\delta_{\dot{\mathbf{x}}'_1}(\dot{\mathbf{x}}_q)$ is a point mass located at $\dot{\mathbf{x}}_1$. As the posterior continues to be Dirichlet distributed, the posterior is also a Dirichlet process with concentration parameter that simply adds the count of new samples to the previous count and a base probability that is the proportion of the total number of samples in which say state $\dot{\mathbf{x}}_1$ occurs. As such, the Dirichlet process provides a conjugate family of priors over distributions.

By modeling the dynamic uncertainty over the transition function of distinguished state factors as a Dirichlet process, we may limit our attention to the counts of the different state samples. Let ϕ be the vector of counts of all transitions; its size is $|\dot{\mathbf{X}}|^2|A|$. Next, we show how to include the Dirichlet process in I-POMDP-Lite.

We augment the state space of I-POMDP$^{\mathcal{L}}_{i,l}$ to include this vector: $\mathcal{S} = S \times \Phi$ where Φ is the space of all such vectors and is of size $\mathbb{N}^{|\dot{\mathbf{X}}|^2|A|}$. Given the augmented state space, we redefine the transition, observation and reward functions

of I-POMDP$^{\mathcal{L}}_{i,l}$ as follows:

$$\mathcal{T}_i(\langle s,\phi\rangle, a_i, \mathbf{a}_{-i}, \langle s',\phi'\rangle) = \begin{cases} T_i(s_{/\dot{\mathbf{x}}}, a_i, \mathbf{a}_{-i}, s'_{/\dot{\mathbf{x}}}) & \text{if } \phi' = \phi + \delta_{\dot{\mathbf{x}}\mathbf{a}\dot{\mathbf{x}}'} \\ \times E[\dot{T}_i(\dot{\mathbf{x}}, a_i, \mathbf{a}_{-i}, \dot{\mathbf{x}}')] & \\ 0 & \text{otherwise} \end{cases}$$

Here, $\delta_{\dot{\mathbf{x}}\mathbf{a}\dot{\mathbf{x}}'}$ is a vector of size $|\dot{\mathbf{X}}|^2|A|$ with all 0s except for a 1 at the location indexed by $(\dot{\mathbf{x}}, \mathbf{a}, \dot{\mathbf{x}}')$ where $\dot{\mathbf{x}}$ is the distinguished factor of state s and $\dot{\mathbf{x}}'$ is a factor of s'. *The expected transition probability is obtained from the posterior Dirichlet process.* The observation function is now defined as,

$$\mathcal{O}_i(\langle s,\phi\rangle, a_i, \mathbf{a}_{-i}, \langle s',\phi'\rangle, o_i) = \begin{cases} O_i(s', a_i, \mathbf{a}_{-i}, o_i) & \text{if } \phi' = \phi + \delta_{\dot{\mathbf{x}}\mathbf{a}\dot{\mathbf{x}}'} \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

The reward function is straightforward: $\mathcal{R}_i(\langle s,\phi\rangle, a_i, \mathbf{a}_{-i}) = R_i(s, a_i, \mathbf{a}_{-i})$. The optimality criteria remains the same as before.

Consequently, the *augmented* I-POMDP$^{\mathcal{L}}_{i,l}$ is defined by the tuple $\langle \mathcal{S}, A, \mathcal{T}_i, \Omega_i, \mathcal{O}_i, \mathcal{R}_i, \{\mathcal{M}_{j,l-1}, \mathcal{M}_{k,l-1}, \ldots, \mathcal{M}_{z,l-1}\}, OC_i\rangle$, where the new parameters are defined as above. It shares commonality with the Bayes-adaptive I-POMDP framework [16] though we are uncertain over partial transition distributions only and our framework differs by limiting attention to nested MDPs as models of others. Acting optimally in response to observations in this augmented framework entails the standard balance between exploring to learn the transition distributions of the distinguished state factor(s) with greater confidence and exploiting the learned distributions for reward. However, compared to traditional online methods for reinforcement learning, this balance is achieved offline as an integral part of the planning.

The exact solution of the augmented I-POMDP$^{\mathcal{L}}_{i,l}$ is challenged by the infinite state space because the count vector ϕ grows unboundedly. If the count vector somehow reflects the true transition probabilities, then $\mathcal{T}_i$ effectively collapses into the true transition function and we may obtain the exact solution of the planning problem. However, by the law of large numbers we can only approach the true distributions asymptotically using counts. Nevertheless, the following observation provides guidance on how we can move forward:

Observation 3. *With increasing numbers of samples, means of the posterior Dirichlet processes $DP(n, C)$ come arbitrarily close to the true transition probabilities. Consequently, values of the policies using the estimated transition functions may also come arbitrarily close to the value of the exact policy.*

Indeed, Ross et al. [17] exploit the above observation in the context of POMDPs and identify an ϵ-dependent finite space of counts of both transitions and observations whose consideration leads to policies with values that are within ϵ of the exact (obtained using the infinite space). We extend these results to our context where the uncertainty is over the partial transition function only but involving multiple agents; this allows solving the augmented I-POMDP$^{\mathcal{L}}_{i,l}$ with finite state spaces as bounded approximation of the exact.

Let $\alpha^t(s,\phi;\pi_{i,l})$ be i's expected value of following policy $\pi_{i,l}$ from augmented state (s,ϕ) at some time step t. Let $\mathcal{N}^{\dot{\mathbf{x}}\mathbf{a}}_{\phi} = \sum_{q=1}^{|\dot{X}|} \phi_{\dot{\mathbf{x}}\mathbf{a}\dot{\mathbf{x}}_q}$ where $\phi_{\dot{\mathbf{x}}\mathbf{a}\dot{\mathbf{x}}_q}$ is the count for the transition $(\dot{\mathbf{x}}, \mathbf{a}, \dot{\mathbf{x}}_q)$ contained in the vector ϕ; and $\mathcal{N}^{\epsilon} = \max\left(\frac{|\dot{\mathbf{X}}|(1+\epsilon')}{\epsilon'}, \frac{1}{\epsilon''} - 1\right)$ where $\epsilon' = \frac{\epsilon(1-\gamma)^2}{8\gamma\mathcal{R}_{max}}$ and $\epsilon'' = \frac{\epsilon(1-\gamma)^2\, ln(\gamma^{-\epsilon})}{32\gamma\mathcal{R}_{max}}$. Here, $\mathcal{R}_{max}$ is the largest value in $\mathcal{R}$.

Proposition 1 shows that for transition counts that exceed $\mathcal{N}^{\epsilon}$, there exist counts less than or equal to $\mathcal{N}^{\epsilon}$ such that the negative impact of the reduced count on the expected value of following policy $\pi_{i,l}$ from the same state is bounded. More formally,

Proposition 1 (Bounded difference in value)**.** *Given $\epsilon > 0$ and for any (s,ϕ) such that $\mathcal{N}^{\dot{\mathbf{x}}\mathbf{a}}_{\phi} > \mathcal{N}^{\epsilon}$ for all $\dot{\mathbf{x}}, \mathbf{a}$, there exist ϕ' such that $\mathcal{N}^{\dot{\mathbf{x}}\mathbf{a}}_{\phi'} \leq \mathcal{N}^{\epsilon}$ for all $\dot{\mathbf{x}}, \mathbf{a}$, and $|\alpha^h(s,\phi;\pi_{i,l}) - \alpha^h(s,\phi';\pi_{i,l})| \leq \epsilon$.*

The proof of this proposition extends the proof of a similar proposition by Ross et al. [17] to the multiagent context of I-POMDP$^{\mathcal{L}}_{i,l}$ in a straightforward way. Let $\mathcal{S}^{\epsilon}$ be the set of augmented states of I-POMDP$^{\mathcal{L}}_{i,l}$ such that the count vectors are all limited to the following set, $\Phi^{\epsilon} = \{\phi \in \Phi : \mathcal{N}^{\dot{\mathbf{x}}\mathbf{a}}_{\phi} \leq \mathcal{N}^{\epsilon} \ \forall \dot{\mathbf{x}}, \mathbf{a}\}$; in other words, $\mathcal{S}^{\epsilon} = \mathcal{S} \times \Phi^{\epsilon}$. Then, define a new transition function over the augmented and bounded state space, $\mathcal{T}^{\epsilon}_i : \mathcal{S}^{\epsilon} \times A \times \mathcal{S}^{\epsilon} \to [0,1]$ such that

$$\mathcal{T}^{\epsilon}_i(\langle s,\phi\rangle, a_i, \mathbf{a}_{-i}, \langle s',\phi'\rangle) = \begin{cases} T_i(s_{/\dot{\mathbf{x}}}, a_i, \mathbf{a}_{-i}, s'_{/\dot{\mathbf{x}}}) & \text{if } \phi' = \zeta(\phi + \delta_{\dot{\mathbf{x}}\mathbf{a}\dot{\mathbf{x}}'}) \\ \times E[\dot{T}_i(\dot{\mathbf{x}}, a_i, \mathbf{a}_{-i}, \dot{\mathbf{x}}')] & \\ 0 & \text{otherwise} \end{cases}$$

Here, ζ is a function that projects those counts which cause $\mathcal{N}^{\dot{\mathbf{x}}\mathbf{a}}_{\phi}$ to exceed $\mathcal{N}^{\epsilon}$ back to values so that the latter is not exceeded. If $\phi + \delta_{\dot{\mathbf{x}}\mathbf{a}\cdot}$ does not exceed $\mathcal{N}^{\epsilon}$, then ζ is an identity function. Observation function with the bounded state space also applies the projection ζ to Eq. 7 similarly to its use above. Finally, the reward function $\mathcal{R}^{\epsilon}_i(\langle s,\phi\rangle, a_i, \mathbf{a}_{-i}) = R_i(s, a_i, \mathbf{a}_{-i})$.

Subsequently, the definition of I-POMDP$^{\mathcal{L}}_{i,l}$ modifies to be the tuple $\langle \mathcal{S}^{\epsilon}, A, \mathcal{T}^{\epsilon}_i, \Omega_i, O^{\epsilon}_i, \mathcal{R}_i, \{\mathcal{M}_{j,l-1}, \mathcal{M}_{k,l-1}, \ldots, \mathcal{M}_{z,l-1}\}, OC_i\rangle$. Let $\alpha^{h,\epsilon}(s,\phi;\pi_{i,l})$ be the expected value of following policy $\pi_{i,l}$ from state (s,ϕ) according to the modified framework with the bounded state space. Then, we may bound the negative impact on expected value due to using the new framework as follows:

Proposition 2 (Bounded difference in convergent value)**.** *Given $\epsilon > 0$ and the augmented I-POMDP$^{\mathcal{L},\epsilon}_{i,l}$ with the bounded state space $\mathcal{S}^{\epsilon}$, the following holds:*

$$|\alpha^{h}(s,\phi;\pi_{i,l}) - \alpha^{h,\epsilon}(s,\phi';\pi_{i,l})| \leq \frac{\epsilon}{1-\gamma}$$

for any $(s,\phi) \in \mathcal{S}$ and some $(s,\phi') \in \mathcal{S}^{\epsilon}$ where $\phi' = \zeta(\phi)$.

The proof of this proposition essentially generalizes Prop. 1 to the infinite horizon and is given in the Appendix. Consequently, Prop. 2 allows us to solve the augmented I-POMDP$^{\mathcal{L},\epsilon}_{i,l}$ while incurring a bounded loss.

In the context of open agent systems, the augmented framework provides a way to learn the transition probabilities of state factors that influence the other agents' decisions about whether they ought to leave the network, as a part of planning.

5 EXPERIMENTS

While the predictive method has the obvious advantage of potentially anticipating agent departures, it must first learn the transition probabilities accurately. Consequently, an empirical evaluation of the presented approaches on multiple configurations is needed.

5.1 Setup

We empirically evaluate our methods labeled **I-PBVI PostHoc** and **I-PBVI Predictive** using a realistic simulation of the complex wildfire domain (adapted from [18], similar to [19]). In the simulation, an agent obtains a reward of 1 each step and for each location that is not on fire, and a penalty of 100 for doing anything but a NOOP while recharging suppressant or trying to fight a nonexistent fire. Agents have three suppressant levels: empty and recharging, half full, and full with stochastic transitions between levels.

We measure the performance of agents employing both methods in two ways: (1) the average of discounted and cumulative rewards obtained by each agent; and (2) the average intensity of each fire over time (where intensity ranges from 0 for no fire to 4 for a burned-out location). The former evaluates the agent's planning method for maximizing rewards, whereas the latter evaluates the system-level performance of the team of agents in achieving their overall objective – suppressing the wildfire in the forest. Furthermore, we also include the performance of baselines that represent what would happen to the forest (1) if no agents were present (called **NOOP** as this situation is equivalent to all agents always taking NOOP actions), (2) if each agent randomly chooses between actions that put out fires or NOOP (labeled **Random**), representing a scenario where agents do not plan how or when to interact with their peers, and (3) if each agent carries out actions selected according to a heuristic-variant of Random (called **Heuristic**) – fight existing fires (chosen by random selection) only if the agent has available suppressant, else take a NOOP.

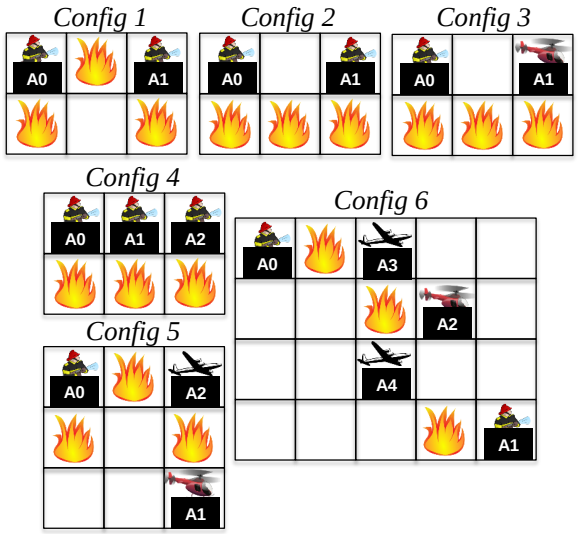


Figure 2: Illustration of experimental configurations.

To elicit different interactions between agents, we consider six configurations in our experiments, illustrated in Fig. 2: configurations C1, C2, and C3 where two agents are responsible for protecting the forest with three fires in each 2×3 grid, configuration C4 where three agents fight three fires in a 2×3 grid, configuration C5 where three agents fight three fires in a more spread out 3×3 grid, and configuration 6 where five agents fight 3 fires in a much larger 5×4 grid. In each of the six configurations, an agent can only put out a fire that is immediately adjacent – to its south, north, east, or west, or diagonal from the agent. In each configuration, I-PBVI PostHoc agents assume a uniform transition distribution for how peers' suppressant levels change, whereas I-PBVI Predictive agents perform random actions in simulation to learn the transition dynamics to better model openness and its impacts on joint behavior. After learning a transition model (i.e., after 100 steps and 30 trials), each agent in the predictive method will use this model for planning.

First configuration C1 contains two agents, each with an individual fire to fight (F0, F2, respectively) while also sharing a fire (F1). Each agent in C1 can lower the intensity of a fire by one. Configuration C2 represents an environment similar to C1, except all three fires are adjacent, and thus can spread to neighboring locations, increasing the pressure on agents to control the wildfires in the environment. Configuration C3, on the other hand, is the same as C2 except that agents are of differing types: A0 lowers the intensity of a fire by one, while A1 is more powerful and can accomplish twice as much reduction when it fights a fire. Together,

Table 1: Average team discounted rewards with 95% confidence intervals.

Configuration	I-PBVI Predictive	I-PBVI PostHoc	Heuristic	Random	NOOP
C1	16.635 ± 1.613	15.001 ± 0.479	5.709 ± 1.952	-380.114 ± 56.552	0.000 ± 0.000
C2	14.706 ± 0.573	14.681 ± 0.422	7.517 ± 2.369	-442.186 ± 58.427	0.000 ± 0.000
C3	27.459 ± 1.107	26.202 ± 1.267	11.455 ± 2.811	-488.988 ± 66.531	0.000 ± 0.000
C4	$\mathbf{49.607 \pm 3.211}$	44.340 ± 3.235	26.419 ± 3.606	-552.740 ± 74.620	0.000 ± 0.000
C5	$\mathbf{49.341 \pm 0.859}$	48.676 ± 1.773	25.365 ± 2.052	-844.951 ± 50.143	0.000 ± 0.000
C6	$\mathbf{103.735 \pm 2.859}$	87.532 ± 1.432	56.006 ± 14.481	-1272.801 ± 37.768	0.000 ± 0.000

Table 2: Average fire intensities with 95% confidence intervals.

Configuration	I-PBVI Predictive	I-PBVI PostHoc	Heuristic	Random	NOOP
C1	$\mathbf{2.597 \pm 0.038}$	2.686 ± 0.038	3.063 ± 0.032	3.379 ± 0.027	3.948 ± 0.005
C2	2.683 ± 0.038	2.695 ± 0.038	3.053 ± 0.033	3.630 ± 0.021	3.953 ± 0.004
C3	$\mathbf{1.537 \pm 0.039}$	1.670 ± 0.039	2.806 ± 0.035	3.250 ± 0.030	3.953 ± 0.004
C4	$\mathbf{0.834 \pm 0.031}$	1.024 ± 0.034	2.068 ± 0.039	2.841 ± 0.035	3.954 ± 0.004
C5	$\mathbf{1.361 \pm 0.038}$	1.374 ± 0.038	1.929 ± 0.040	2.222 ± 0.040	3.953 ± 0.004
C6	$\mathbf{1.025 \pm 0.018}$	1.222 ± 0.019	1.684 ± 0.015	1.999 ± 0.017	3.958 ± 0.002

these configurations enable us to evaluate (1) how agents are able to *balance between fighting a shared fire and their own individual fires* in C1; (2) how agents *behave under a more pressing situation* in C2; and (3) how *different types of agents interact* in C3.

In configuration C4, we extend C2 to add a third agent, which simultaneously *makes others' actions more difficult to predict* since each agent has an extra neighbor that it can work together with, while at the same time *provides more firefighting ability* to control wildfires in the forest. Note that agent A1 in C4 can fight all 3 fires. Configuration C5 adds to the complexity – it not only represents a larger, more spread out forest but it also involves *more intricate relationships among three agents*, each of different types. Namely, the three agents A0, A1, and A2 can each reduce the intensity of a fire by 1, 2, and 3 with each firefighting action, respectively. Thus, A2 (who shares fires with both A0 and A1) is quite powerful and is able to put out fires entirely by itself. As a result, its neighbors face interesting decisions of predicting what A2 will do in order to choose their optimal best response (either fighting a different fire, or conserving suppressant to fight future fires). Configuration C6 further adds to the complexity of C5 – it represents a much larger forest involving five agents. Agents A0-4 in C6 can reduce the fire intensity by 1, 1, 2, 3, and 3 respectively. Comparing all six of these configurations, we note that the complexity of agent reasoning increases as the configuration number increases, because more fire locations are shared between agents, more agents interact with one another, and more types of agents are introduced in the environment. For each configuration, we conducted 30 runs of 100 steps, and we average the results of our performance measures.

5.2 Results and Analysis

Tables 1 and 2 present the average discounted, cumulative rewards earned (summed across the team of agents) and the average intensity across all fires per time step, respectively. From these results, we first observe that agents using the I-PBVI Predictive and PostHoc solutions earned greater cumulative rewards, as well as achieved lower average fire intensities than the baseline approaches. This indicates that our approaches to planning about the presence of peer agents in open environments is indeed beneficial toward both agent performance as measured by cumulative rewards, as well as desired system behavior due to reduction in wildfires.

Comparing between our two approaches, we make several additional important observations. First, I-PBVI Predictive performed better than I-PBVI PostHoc in all configurations in terms of average fire intensity (with statistical significance at the $p = 0.05$ level in configurations C1, C3, C4, C5 and C6). Thus, learning how to predict when peer agents will be available and when they might be absent from the environment is indeed beneficial to helping agents achieve system-level goals (i.e., minimizing fires in our domain). In terms of discounted rewards, I-PBVI Predictive also outperformed I-PBVI PostHoc in larger configurations but with only a slight (non-statistically significant) advantage in smaller ones.

To better understand the differences in agent behavior produced by I-PBVI Predictive and PostHoc, we further investigated the different types of interactions between agents. These interactions are based on the types of actions chosen by agents, including putting out individual fires, collaborating with another agent in fighting a fire, fighting alone a fire that is shared by multiple agents, performing a NOOP due to recharging the agent's suppressant, performing a NOOP because there was no fire to fight, and performing a NOOP to conserve suppressant instead of fighting an available fire.

We discovered that I-PBVI Predictive consistently carried out a higher percentage of NOOP actions in order to conserve suppressant than did I-PBVI PostHoc–1.23 to 1.56 times more in configurations C1, C2, and C3, 2.18 times more in configuration C4, 20.36 times more in configuration

C5 and 28.21 times more in C6. Thus, I-PBVI Predictive caused an agent to conserve its valuable, limited suppressant so that it would be able to contribute when its potential partner agent becomes available to jointly fight the shared fire (as indicated by the combination of more NOOPs when fires were present and lower overall average fire intensity). Further, this provides evidence that learning to predict the presence of neighbors in open environments (using I-PBVI Predictive) does lead to agents that better consider the impacts of interactions between their joint actions, which in turn results in better global behavior toward system goals.

We also discovered that I-PBVI Predictive caused agents to be 2.85 and 1.13 times more likely to fight their own individual fires in C1 and C3, respectively, when there were fewest agents available to fight fires and more individual behavior was necessary. In the similar C2 environment where fires spread faster than C1 and agents had less overall firefighting ability than in C3 (where one agent could reduce fires faster), and thus it was more difficult to fight individual fires than in C1 and C3, both I-PBVI Predictive and PostHoc focused solely on fighting the joint fire that they could feasibly extinguish together. These results further indicate that learning to predict the presence of peers also helps agents better balance individual-centered behavior vs. collaborative behavior in open environments, depending on the needs of the environment.

6 CONCLUSION

As a first paper on modeling open agent systems from a decision-theoretic perspective, the focus of this effort was to study the impact of agents leaving and reentering from the perspective of an individual agent and to point out areas where existing frameworks can be generalized to tackle this problem in a principled manner. As an immediate next step, we are looking into Monte-Carlo based approaches for better scalability during planning. Furthermore, we are currently exploring how *anonymity* – the problem structure that it doesn't matter who fights the fire, but how many agents fight it – can be featured into frameworks like I-POMDP-Lite. Anonymity coupled with better planners may help scale to real-world problems involving 1000+ agents.

ACKNOWLEDGMENTS

This research is supported in part by grants from NSF IIS-0845036 and a grant from ONR N-00-0-141310870 (to PD). We thank James MacGlashan who authored the BURLAP codebase for invaluable assistance during implementation.

Appendix

Proposition 1 (Bounded difference in convergent value)**.** *Given $\epsilon > 0$ and the augmented I-POMDP$^{\mathcal{L}}_{i,l}$ with the bounded state space $\mathcal{S}^{\epsilon}$, the following holds:*

$$|\alpha^{h,\epsilon}(s,\phi';\pi_{i,l}) - \alpha^{h}(s,\phi;\pi_{i,l})| \leq \frac{\epsilon}{1-\gamma}$$

for any $(s,\phi) \in \mathcal{S}$ and some $(s,\phi') \in \mathcal{S}^{\epsilon}$ where $\phi' = \zeta(\phi)$.

Proof. The proof of this proposition essentially generalizes Prop. 1 to the infinite horizon. Let $\mathcal{E}^h = \max_{\alpha^{h,\epsilon},\alpha^h,s,\phi} |\alpha^{h,\epsilon}(s,\phi';\pi_{i,l}) - \alpha^h(s,\phi;\pi_{i,l})|$. Omitting writing the $\max$ norm for brevity, we have,

$$\begin{aligned}
\mathcal{E} &= |\alpha^{h,\epsilon}(s,\phi';\pi_{i,l}) - \alpha^{h}(s,\phi';\pi_{i,l}) + \alpha^{h}(s,\phi';\pi_{i,l}) \\
&- \alpha^{h}(s,\phi;\pi_{i,l})| \\
&\leq |\alpha^{h,\epsilon}(s,\phi';\pi_{i,l}) - \alpha^{h}(s,\phi';\pi_{i,l})| + |\alpha^{h}(s,\phi';\pi_{i,l}) \\
&- \alpha^{h}(s,\phi;\pi_{i,l})| \text{(from the law of triangle inequality)} \\
&\leq |\alpha^{h,\epsilon}(s,\phi';\pi_{i,l}) - \alpha^{h}(s,\phi';\pi_{i,l})| + \epsilon \quad \text{(from Prop. 1)} \\
&= |\mathcal{R}^{\epsilon}_i(s,\phi',a_i,\mathbf{a}_{-i}) + \gamma \sum_{s'\in S, o_i\in\Omega_i} \mathcal{T}^{\epsilon}_i(\langle s,\phi'\rangle, a_i, \mathbf{a}_{-i}, \\
&\langle s',\phi''\rangle) \times \mathcal{O}^{\epsilon}_i(\langle s',\phi''\rangle, a_i, \mathbf{a}_{-i}, o_i)\ \alpha^{h-1,\epsilon}_{o_i}(s',\phi';\pi_{i,l}) \\
&- \mathcal{R}_i(s,\phi',a_i,\mathbf{a}_{-i}) + \gamma \sum_{s'\in S, o_i\in\Omega_i} \mathcal{T}_i(\langle s,\phi'\rangle, a_i, \mathbf{a}_{-i}, \langle s',\phi''\rangle) \\
&\mathcal{O}_i(\langle s',\phi''\rangle, a_i, \mathbf{a}_{-i}, o_i) \times \alpha^{h-1}_{o_i}(s',\phi';\pi_{i,l}1)| + \epsilon \\
&= |R_i(s,\phi' a_i,\mathbf{a}_{-i}) + \gamma \sum_{s'\in S, o_i\in\Omega_i} T_i(s_{/\dot{\mathbf{x}}}, a_i, \mathbf{a}_{-i}, s'_{/\dot{\mathbf{x}}}\rangle) \\
&E[\dot{T}_i(\dot{\mathbf{x}}, a_i, \mathbf{a}_{-i}, \dot{\mathbf{x}}')]\ O_i(s', a_i, \mathbf{a}_{-i}, o_i)\ \alpha^{h-1,\epsilon}_{o_i}(s',\phi';\pi_{i,l}) \\
&- R_i(s,\phi',a_i,\mathbf{a}_{-i}) + \gamma \sum_{s'\in S, o_i\in\Omega_i} T_i(s_{/\dot{\mathbf{x}}}, a_i, \mathbf{a}_{-i}, s'_{/\dot{\mathbf{x}}}) \\
&E[\dot{T}_i(\dot{\mathbf{x}}, a_i, \mathbf{a}_{-i}, \dot{\mathbf{x}}')] \times O_i(s', a_i, \mathbf{a}_{-i}, o_i) \\
&\alpha^{h-1}_{o_i}(s',\phi';\pi_{i,l}1)| + \epsilon \\
&\leq \gamma \sum_{s'\in S, o_i\in\Omega_i} T_i(s_{/\dot{\mathbf{x}}}, a_i, \mathbf{a}_{-i}, s'_{/\dot{\mathbf{x}}}\rangle)\ E[\dot{T}_i(\dot{\mathbf{x}}, a_i, \mathbf{a}_{-i}, \dot{\mathbf{x}}')] \\
&O_i(s', a_i, \mathbf{a}_{-i}, o_i)|\alpha^{h-1,\epsilon}_{o_i}(s',\phi'';\pi_{i,l}) - \alpha^{h-1}_{o_i}(s',\phi'';\pi_{i,l})| + \epsilon \\
&\leq \gamma \sum_{s'\in S, o_i\in\Omega_i} T_i(s_{/\dot{\mathbf{x}}}, a_i, \mathbf{a}_{-i}, s'_{/\dot{\mathbf{x}}}\rangle)\ E[\dot{T}_i(\dot{\mathbf{x}}, a_i, \mathbf{a}_{-i}, \dot{\mathbf{x}}')] \\
&O_i(s', a_i, \mathbf{a}_{-i}, o_i) \max_{s',o_i,\phi''} |\alpha^{h+1,\epsilon}_{o_i}(s',\phi'';\pi_{i,l}) \\
&- \alpha^{h+1}_{o_i}(s',\phi'';\pi_{i,l})| + \epsilon \\
&= \gamma \max_{s',o_i,\phi''} |\alpha^{h-1,\epsilon}(s',\phi'';\pi_{i,l}) - \alpha^{h-1}(s',\phi'';\pi_{i,l})| + \epsilon \\
&= \gamma\, \mathcal{E}^{h-1} + \epsilon
\end{aligned}$$

Notice that $|\alpha^{1,\epsilon}(s',\phi'';\pi_{i,l}) - \alpha^{1}(s',\phi'';\pi_{i,l})| = |\mathcal{R}^{\epsilon}_i(s',\phi'',a_i,\mathbf{a}_{-i}) - \mathcal{R}_i(s',\phi'',a_i,\mathbf{a}_{-i})| = |R_i(s',a_i,\mathbf{a}_{-i}) - R_i(s',a_i,\mathbf{a}_{-i})| = 0$. The above recursion is a geometric series with a base case of 0. Therefore, $\mathcal{E}^h \leq \frac{\epsilon}{1-\gamma}$. □

References

[1] O. Shehory. Software architecture attributes of multi-agent systems. In *1st International Workshop on Agent-Oriented Software Engineering, Revised papers*, pages 77–89. ACM, 2000.

[2] Piotr J. Gmytrasiewicz and Prashant Doshi. A framework for sequential planning in multiagent settings. *Journal of Artificial Intelligence Research*, 24:49–79, 2005.

[3] Prashant Doshi. Decision making in complex multiagent settings: A tale of two frameworks. *AI Magazine*, 33(4):82–95, 2012.

[4] Daniel S. Bernstein, Robert Givan, Neil Immerman, and Shlomo Zilberstein. The complexity of decentralized control of Markov decision processes. *Mathematics of Operations Research*, 27(4):819–840, 2002.

[5] Trong Nghia Hoang and Kian Hsiang Low. Interactive POMDP lite: Towards practical planning to predict and exploit intentions for interacting with self-interested agents. In *23rd International Joint Conference on AI (IJCAI)*, 2013.

[6] J. Calmet, A. Daemi, R. Endsuleit, and T. Mie. A liberal approach to openness in societies of agents. In *Engineering Societies in the Agents World IV, Lecture Notes in Computer Science*, volume 3071, pages 81–92, 2004.

[7] W. Jamroga, A. Meski, and M. Szreter. Modularity and openness in modeling multi-agent systems. In *Fourth International Symposium and Games, Automata, Logics and Formal Verification (GandALF)*, pages 224–239, 2013.

[8] J. Jumadinova, P. Dasgupta, , and L.-K. Soh. Strategic capability-learning for improved multi-agent collaboration in ad-hoc environments. *IEEE Transactions on Systems, Man, and Cybernetics-Part A*, 44(8):1003–1014, 2014.

[9] B. Chen, X. Chen, A. Timsina, and L.-K. Soh. Considering agent and task openness in ad hoc team formation (extended abstract). In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 1861–1862, 2015.

[10] T. D. Huynh, N. R. Jennings an, and N. R. Shadbolt. An integrated trust and reputation model for open multi-agent systems. *Journal of Autonomous Agents and Multi-Agent Systems*, 13(2):119–154, 2006.

[11] I. Pinyol and J. Sabater-Mir. Computational trust and reputation models for open multi-agent systems: A review. *Artificial Intelligence Review*, 40:1–25, 2011.

[12] Peter Stone, Gal A. Kaminka, Sarit Kraus, and Jeffrey S. Rosenschein. Ad hoc autonomous agent teams: Collaboration without pre-coordination. In *Twenty-Fourth AAAI Conference on Artificial Intelligence*, 2010.

[13] Ranjit Nair, Pradeep Varakantham, Milind Tambe, and Makoto Yokoo. Networked distributed POMDPs: A synthesis of distributed constraint optimization and POMDPs. In *Twentieth AAAI Conference on Artificial Intelligence*, pages 133–139, 2005.

[14] Yoonheui Kim, Ranjit Nair, Pradeep Varakantham, Milind Tambe, and Makoto Yokoo. Exploiting locality of interaction in networked distributed POMDPs. In *AAAI Spring Symposium on Distributed Plan and Schedule Management*, 2006.

[15] Frans Oliehoek, Matthijs Spaan, Shimon Whiteson, and Nikos Vlassis. Exploiting locality of interaction in factored dec-POMDPs. In *Seventh International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 517–524, 2008.

[16] Brenda Ng, Kofi Boakye, Carol Meyers, and Andrew Wang. Bayes-adaptive interactive POMDPs. In *Twenty-Sixth AAAI Conference on Artificial Intelligence (AAAI)*, pages 1408 – 1414, 2012.

[17] Stephane Ross, Brahim Chaib-draa, and Joelle Pineau. Bayes-adaptive POMDPs. In *Neural Information Processing Systems (NIPS)*, 2007.

[18] D. Boychuk, W.J. Braun, R.J. Kulperger, Z.L. Krougly, and D.A. Stanford. A stochastic forest fire growth model. *Environmental and Ecological Statistics*, 16(2):133–151, 2009.

[19] N.K. Ure, S. Omidshafiei, B.T. Lopez, A.-A. Agha-Mohammadi, J.P. How, and J. Vian. Online heterogeneous multiagent learning under limited communication with applications to forest fire management. In *Intelligent Robotics and Systems (IROS)*, pages 5181–5188, 2015.