MULTI-AGENT DECENTRALIZED BELIEF PROPAGATION ON GRAPHS

Anonymous authors
Paper under double-blind review

ABSTRACT

We consider the problem of interactive partially observable Markov decision processes (I-POMDPs), where the agents are located at the nodes of a communication network. Specifically, we assume a certain message type for all messages. Moreover, each agent makes individual decisions based on the interactive belief states, the information observed locally and the messages received from its neighbors over the network.

Within this setting, the collective goal of the agents is to maximize the globally averaged return over the network through exchanging information with their neighbors. We propose a decentralized belief propagation algorithm for the problem, and prove the convergence of our algorithm. Finally we show multiple applications of our framework. Our work appears to be the first study of decentralized belief propagation algorithm for networked multi-agent I-POMDPs.

1 INTRODUCTION

In reinforcement learning, Partially Observable Markov Decision Processes (POMDPs) (KLC98; BDH99; RN02) is a general decision-theoretic framework for planning under uncertainty in a partially observable, stochastic environment. An agent makes decisions rationally in such settings by maintaining beliefs of the physical state and sequentially choosing the optimal actions that maximize the expected value of future rewards. Solutions of POMDPs are mappings from an agent's beliefs to actions. The drawback of POMDPs in the multi-agent scenario is that the impact of other agents' actions cannot be represented explicitly. Examples of such POMDPs are infinite generalized policy representation (LLC11), and infinite POMDPs (DVPWR13).

Interactive POMDPs (I-POMDPs) (GD05) generalize POMDPs to multi-agent settings by substituting POMDP belief spaces with interactive belief spaces. More specifically, an I-POMDP substitutes plain beliefs of the state space with augmented beliefs of the state space and the other agents' beliefs and models. The models of other agents included in the new augmented belief space consist of two types: intentional models and subintentional models. An intentional model ascribes beliefs, preferences, and rationality to other agents, while a simpler subintentional model, such as finite state controllers (PG16) does not. The augmentation of intentional models forms a hierarchical belief structure that represents an agent's beliefs of the physical state, beliefs of the other agents and their beliefs of others' beliefs, and it can be nested to arbitrary levels. Solutions of I-POMDPs map an agent's belief of the environment and other agents' models to actions. It has been shown that the augmented belief in I-POMDP results in a higher value function compared to one obtained from POMDP, which implies I-POMDPs' modeling superiority.

In this work, we study the problem of POMDPs with collaborative agents. For collaborative POMDPs problem, it is important to specify the interactions bewteen the agents. One appealing option is to have a central controller which has the information of all agents, and determines the actions for all agents. With all the information available to the controller, the problem reduces to a classical POMDP and can be solved by existing single-agent POMDP algorithms. Yet, in a bunch of real-world applications, such as sensor networks (ASSC02; RN04) and intelligent transportation systems (AB02), it may be very costly to have a central controller. Moreover, since the central controller needs to communicate with each agent to get information, it increases the communication overhead at each controller. The communication overhead degrades the scalability of the multi-agent system as well as its robustness to malicious attacks. Another option is to make

the collaboration between agents implicitly represented by the I-POMDP model. Although agents' actions do not change the other agents' model directly, they can change the other agents' belief states indirectly, typically by changing the environment in a way observable to the other agents. The influence to the other agents' belief states can be viewed as some form of exchange of information. However, the modeling superiority of I-POMDPs comes at the cost of a drastic increase of the belief space complexity, because the agent models grow exponentially as the belief nesting level increases. Hence, the complexity of the belief representation is proportional to belief dimensions, which is known as the curse of dimensionality. Moreover, due to the fact that exact solutions to POMDPs are PSPACE-complete and undecidable for finite and infinite time horizon respectively (PT87), the time complexity of more generalized I-POMDPs is at least PSPACE-complete and undecidable for finite and infinite horizon, since an I-POMDPs may contain multiple POMDPs or I-POMDPs of other agents.

Given all these disadvantages of the centralized model and the complexity of I-POMDPs, we consider a decentralized model where the agents in I-POMDPs are connected by a communication network. Specifically, let $\{\mathcal{G} = (\mathcal{V}, \mathcal{E})\}$ be a communication network, where $\mathcal{V}$ is the set of nodes, and $\mathcal{E} \subseteq \{(i, j) : i, j \in \mathcal{V}\}$ is the set of edges. We assume that each node represents an agent. And agent $i \in \mathcal{V}$ and agent $j \in \mathcal{V}$ can communicate with each other if and only if $(i, j) \in \mathcal{V}$. As such, at each time slot t, each agent executes an individual action based on the interactive belief states, the local information and the messages sent from its neighbors, with the goal of maximizing the individual average rewards. The message type is crucial in reducing the complexity of I-POMDPs, for example, if the message is a belief, the agent does not have to maintain complex and infinitely nested beliefs of other agents. We call this model as *networked multi-agent I-POMDPs*, which is presented in Section 3 in detail.

Main Contribution. Our contribution in this work is three-fold. First, we formulate interactive POMDPs problem for networked agents, and prove a version of belief update and value iteration update adapted to this setting. Second, we propose a decentralized belief propagation algorithm, and prove the convergence of the proposed algorithm. Third, we show our framework precisely captures the collaboration in decentralized multi-agent cooperative systems by showing multiple applications.

Related Work. AI literature (BGIZ02; NTY$^+$03) appeared in a series of studies that extended POMDP to several branches. One of the branches is called decentralized POMDP (DEC-POMDP), which is related to decentralized control problems (OW96). DEC-POMDP framework assumes that the agents have the common reward function. Furthermore, it is assumed that the optimal joint solution is computed in a central coordinator and then distributed among the agents for execution. (NMT13) shows a variant of DEC-POMDP with a partial historical shared information structure. The framework of I-POMDPs is introduced in (GD05), followed by Bayesian inference approximate solutions (HG18). Another branch of extending POMDPs to multiple agents is called multi-agent POMDP (MPOMDP) (MSL11; AO15). The MPOMDP framework assumes that agents have joint observations, so it can be simplified to POMDP (PT02) by having a single centralized controller that takes joint actions and receives joint observations. In other words, a DEC-POMDP can be reduced to an MPOMDP in a fully communicative scenario.

From the game-theoretic side, most existing works are based on the framework of Markov games, which was first proposed by (Lit94), and then followed by (Lit01; LR00; HW03). This framework applies to the setting with both collaborative and competitive relationships among agents. More recently, several multi-agent reinforcement learning (MARL) algorithms using deep neural networks as function approximators have gained increasing attention (FADFW16; GEK17; BBX$^+$16; OPA$^+$17; FNF$^+$17). A more relevant work is (ZYL$^+$18), the authors study a MARL framework with networked agents, where the communication among agents contributes toward the overall performance of MARL in a fully decentralized setting.

The remainder of this paper is structured as follows. We start with an overview of partially observable Markov decision processes, the concept of agent types and interactive POMDPs in Section 2. We formulate the networked multi-agent I-POMDPs and present a decentralized belief propagation algorithms for the networked multi-agent I-POMDPs problem in Section 3. We provide theoretical result in Section 4. Several applications of our framework is presented in Section 5. We conclude with a brief summary and some future research directions in Section 6.

2 BACKGROUND

2.1 PARTIALLY OBSERVABLE MARKOV DECISION PROCESS

A partially observable Markov decision process (POMDP) of an agent i is defined as

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

where: S is a set of possible states of the environment. A_i is a set of agent i's actions. T_i is a transition function, i.e., $T_i : S \times A_i \times S \rightarrow [0, 1]$ which describes the dynamics of the environment. Ω_i is the set of agent i's observations. O_i is agent i's observation function, i.e., $O_i : S \times A_i \times \Omega_i \rightarrow [0, 1]$. R_i is the reward function for the agent i, i.e., $R_i : S \times A_i \rightarrow \Re$.

The belief update step of POMDP is shown below:

$$b_i^t(s^t) = \beta O_i(o_i^t, s^t, a_i^{t-1}) \sum_{s^{t-1} \in S} b_i^{t-1}(s^{t-1}) T_i(s^t, a_i^{t-1}, s^{t-1}) \tag{2}$$

where β is the normalizing constant. The belief update step takes into account changes in initial belief b_i^{t-1}, action a_i^{t-1}, and the new observation o_i^t. And $b_i^t(s^t)$ is the new belief for state s^t. It is convenient for us to denote the above belief update step for all states in S as $b_i^t = SE(b_i^{t-1}, a_i^{t-1}, o_i^t)$.

We denote agent i's optimality criterion as OC_i, which specifies how rewards over time are handled. In this work, we concentrate on the infinite horizon criterion with discounting, i.e., the agent maximizes the expected value of the sum of the discounted rewards of an infinite horizon $E(\sum_{t=0}^{\infty} \gamma^t r_t)$, where $0 \leq \gamma \leq 1$ is a discount factor. However, our approach can be easily extended to the other criteria.

The utility associated with a belief state, b_i consists of the maximum immediate rewards due to b_i, together with the discounted expected sum of utilities associated with the updated belief states $SE_i(b_i, a_i, o_i)$:

$$U(b_i) = \max_{a_i \in A_i} \left\{ \sum_{s \in S} b_i(s) R_i(s, a_i) + \gamma \sum_{o_i \in \Omega_i} Pr(o_i | a_i, b_i) U(SE_i(b_i, a_i, o_i)) \right\} \tag{3}$$

And the optimal action, a_i^*, is an element of the set of optimal actions, $OPT(b_i)$, for the belief state b_i, defined as:

$$OPT(b_i) = \arg\max_{a_i \in A_i} \left\{ \sum_{s \in S} b_i(s) R_i(s, a_i) + \gamma \sum_{o_i \in \Omega_i} Pr(o_i | a_i, b_i) U(SE_i(b_i, a_i, o_i)) \right\} \tag{4}$$

2.2 AGENT TYPES AND FRAMES

The following two definitions collect POMDP parameters independent of agent implementation and put them into constructs. The representations are convenient for our analysis, so we list them below,

Definition 1 (Type, (GD05)) *A type of an agent i is, $\theta_i = \langle b_i, A_i, \Omega_i, T_i, O_i, R_i, OC_i \rangle$, where b_i is agent i's state of belief (an element of $\Delta(S)$), OC_i is its optimality criterion, and the rest of the elements are as defined before. Let Θ_i be the set of agent i's types.*

Given type θ_i, and the assumption that the agent is Bayesian-rational, we denote the set of agent's optimal actions as $OPT(\theta_i)$.

Definition 2 (Frame, (GD05)) *A frame of an agent i is, $\hat{\theta}_i = \langle A_i, \Omega_i, T_i, O_i, R_i, OC_i \rangle$. Let $\hat{\Theta}_i$ be the set of agent i's frames.*

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

2.3 INTERACTIVE POMDPS

W.l.o.g., we consider an agent i interacting with only one other agents j.

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

$$\textit{I-POMDP}_{i,l} = \langle IS_{i,l}, A, T_i, \Omega_i, O_i, R_i \rangle \tag{5}$$

where $IS_{i,l}$ is a set of interactive states for agent i, defined as $IS_{i,l} = S \times M_{j,l-1}$, $l \geq 1$. Here M_j is the the set of possible models of agent j, and l is the nesting level. The set of models $M_{j,l-1}$ can be divided into two classes, the intentional models $IM_{j,l-1}$, and subintentional models SM_j. Thus, $M_{j,l-1} = IM_{j,l-1} \cup SM_j$.

The intentional models, $IM_{j,l-1}$, ascribe to the other agent beliefs, preferences and rationality in action selection. Thus they are other agents' types. For example, the intentional model for agent j at level $l-1$ can be defined as $\theta_{j,l-1} = \langle b_{j,l-1}, \hat{\theta}_j \rangle$, where $b_{j,l-1}$ is agent j's belief nested to the level $l-1$, $b_{j,l-1} \in \Delta(IS_{j,l-1})$. We omit the details of subintentional model since it is not a focus of this paper.

The interactive states $IS_{i,l}$ can be defined in an inductive manner:

$$\begin{aligned}
&IS_{i,0} = S \qquad \Theta_{j,0} = \{\langle b_{j,0}, \hat{\theta}_j \rangle : b_{j,0} \in \Delta(IS_{j,0})\} \quad M_{j,0} = \Theta_{j,0} \cup SM_j \\
&IS_{i,1} = S \times M_{j,0} \quad \Theta_{j,1} = \{\langle b_{j,1}, \hat{\theta}_j \rangle : b_{j,1} \in \Delta(IS_{j,1})\} \quad M_{j,1} = \Theta_{j,1} \cup M_{j,0} \\
&\cdots \\
&IS_{i,l} = S \times M_{j,l-1} \quad \Theta_{j,l} = \{\langle b_{j,l}, \hat{\theta}_j \rangle : b_{j,l} \in \Delta(IS_{j,l})\} \quad M_{j,l} = \Theta_{j,l} \cup M_{j,l-1}.
\end{aligned} \tag{6}$$

For the rest of the paper, we omit the level subscription for notation simplicity.

All remaining components in an I-POMDP are similar to those in a POMDP except we also keep the following assumptions in (GD05):

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

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

Next, we define the belief update steps and value iterations for I-POMDPs.

2.3.1 BELIEF UPDATE IN I-POMDPS

The next proposition defines the agent i's belief update function, $b_i^t(is^t) = Pr(is^t | o_i^t, a_i^{t-1}, b_i^{t-1})$, where $is^t \in IS_i$ is an interactive state. We use the belief state estimation function, SE_{θ_i}, as an abbreviation for belief updates for individual states so that

$$b_i^t = SE_{\theta_i}(b_i^{t-1}, a_i^{t-1}, o_i^t). \tag{7}$$

$\tau_{\theta_i}(b_i^{t-1}, a_i^{t-1}, o_i^t, b_i^t)$ will stand for $Pr(b_i^t | b_i^{t-1}, a_i^{t-1}, o_i^t)$. Further below we also define the set of type-dependent optimal actions of an agent, $OPT(\theta_i)$.

Given the definition and assumptions, the interactive belief update can be performed as in the next proposition. [Belief Update] Under the MNM and MNO assumptions, the belief update function for an interactive POMDP $\langle IS_i, A, T_i, \Omega_i, O_i, R_i \rangle$, when m_j in is^t is intentional, is:

$$\begin{aligned}
b_i^t(is^t) = &\beta \sum_{is^{t-1}: \hat{m}_j^{t-1} = \hat{\theta}_j^t} b_i^{t-1}(is^{t-1}) \sum_{a_j^{t-1}} Pr(a_j^{t-1} | \theta_j^{t-1}) O_i(s^t, a^{t-1}, o_i^t) \times T_i(s^{t-1}, a^{t-1}, s^t) \\
&\sum_{o_j^t} \tau_{\theta_j^t}(b_j^{t-1}, a_j^{t-1}, o_j^t, b_j^t) O_j(s^t, a^{t-1}, o_j^t).
\end{aligned} \tag{8}$$

2.3.2 VALUE FUNCTION AND SOLUTION IN I-POMDPS

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

$$U(\theta_i) = \max_{a_i \in A_i} \left\{ \sum_{is} b_i(is) ER_i(is, a_i) + \gamma \sum_{o_i \in \Omega_i} Pr(o_i|a_i, b_i) U(\langle SE_{\theta_i}(b_i, a_i, o_i), \hat{\theta}_i) \right\} \quad (9)$$

where, $ER_i(is, a_i) = \sum_{a_j} R_i(is, a_i, a_j) Pr(a_j|m_j)$. Equation (9) is a basis for value iteration in I-POMDPs.

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

$$OPT(\theta_i) = \arg\max_{a_i \in A_i} \left\{ \sum_{is} b_i(is) ER_i(is, a_i) + \gamma \sum_{o_i \in \Omega_i} Pr(o_i|a_i, b_i) U(\langle SE_{\theta_i}(b_i, a_i, o_i), \hat{\theta}_i) \right\}. \quad (10)$$

3 NETWORKED MULTI-AGENT I-POMDP

Consider a set $\mathcal{V}$ of N agents, labeled by an index $i = 1, 2, ..., N$. Their interaction (communication) is modeled by a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where an edge (i, j) is in $\mathcal{E}$ if and only if agent i interacts (communicates) with agent j. We assume that $\mathcal{G}$ is connected, i.e., there is a path from any node i to any other node j. We denote the set of neighbors of agent i as ∂i. The network induced by interaction and communication could be different, but here we do no distinguish them for simplicity.

Let us define type of message and message before we proceed to define the networked multi-agent I-POMDPs.

Definition 4 (Message Type & Message) *We define the set of message types as $\mathcal{M} = \{\times_{i\in[N]} A_i, \Delta(S), \times_{i\in[N]} \Omega_i\}$, i.e., there are three types of messages, 'action', 'belief state' and 'observation'. A type of message M is an element of the message type set $\mathcal{M}$, i.e., $M \in \mathcal{M}$. A message μ_i^t is an element of message type M, i.e. $\mu_i^t \in M$.*

For example, a message sent by an agent i at time slot t of type 'action' can be represented by $\mu_i^t = a_j^{t-1}$.

Now we can define the networked multi-agent I-POMDP.

Definition 5 (Networked Multi-agent I-POMDP) *A networked multi-agent I-POMDP is characterized by a tuple $\langle \mathcal{G}, \{IS_i\}_{i\in[N]}, \{A_i\}_{i\in[N]}, \{T_i\}_{i\in[N]}, \{\Omega_i\}_{i\in[N]}, \{O_i\}_{i\in[N]}, \{R_i\}_{i\in[N]}, M \rangle$.*

- $\mathcal{G}$ is the communication network.
- For each agent i, the interactive state IS_i is defined as $IS_i = S \times M_{\partial i}$, where ∂i is the set of neighbors of agent i in the communication network $\mathcal{G}$.
- For each agent i, the action space is A_i.
- For each agent i, under the MNM assumption, the transition model is defined as $T_i : S \times A_i \times A_{\partial i} \times S \to [0, 1]$.
- For each agent i, Ω_i is defined as before in the I-POMDP model.
- For each agent i, under the MNO assumption, the observation transition function is defined as $O_i : S \times A_i \times A_{\partial i} \times \Omega_i \to [0, 1]$.
- For each agent i, R_i is defined as $IS_i \times A_i \times A_{\partial i} \to \mathfrak{R}$
- M is the message type.

Similar to I-POMDPs, we can define agent i's belief update function. For simplicity, from now on we assume $\partial i = \{j\}$ (we can extend it to more than one neighbor case easily), the belief update function for agent i is $b_i^t(is^t) = Pr(is^t|o_i^t, a_i^{t-1}, b_i^{t-1}, \mu_j^t)$, where $is^t \in IS_i$ is an interactive state

and μ_j^t is the message sent by an agent i at time slot t. We use the belief state estimation function SE_{θ_i}, as an abbreviation for belief updates for individual states so that

$$b_i^t = SE_{\theta_i}(b_i^{t-1}, a_i^{t-1}, o_i^t, \mu_j^t). \tag{11}$$

The next proposition defines the agent i's belief update function in detail.

[Belief Update] Under the MNM and MNO assumptions, the belief update function for agent i of a networked multi-agent I-POMDP $\langle \mathcal{G}, \{IS_i\}_{i\in[N]}, \{A_i\}_{i\in[N]}, \{T_i\}_{i\in[N]}, \{\Omega_i\}_{i\in[N]}, \{O_i\}_{i\in[N]}, \{R_i\}_{i\in[N]}, M\rangle$, when m_j in is^t is intentional, is:

- When message type is 'action':

$$\begin{aligned} b_i^t(is^t) =& \beta \sum_{is^{t-1}:\hat{m}_j^{t-1}=\hat{\theta}_j^t} b_i^{t-1}(is^{t-1}) O_i(s^t, a^{t-1}, o_i^t) T_i(s^{t-1}, a^{t-1}, s^t) \\ & \sum_{o_j^t} \tau_{\theta_j^t}(b_j^{t-1}, a_j^{t-1}, o_j^t, b_j^t) O_j(s^t, a^{t-1}, o_j^t). \end{aligned} \tag{12}$$

- When message type is 'belief state':

$$\begin{aligned} b_i^t(is^t) =& \beta \sum_{is^{t-1}:\hat{m}_j^{t-1}=\hat{\theta}_j^t} b_i^{t-1}(is^{t-1}) \sum_{a_j^{t-1}\in OPT(\theta_j)} Pr(a_j^{t-1}|\theta_j^{t-1}) \\ & O_i(s^t, a^{t-1}, o_i^t) T_i(s^{t-1}, a^{t-1}, s^t). \end{aligned} \tag{13}$$

- When message type is 'observation':

$$\begin{aligned} b_i^t(is^t) =& \beta \sum_{is^{t-1}:\hat{m}_j^{t-1}=\hat{\theta}_j^t} b_i^{t-1}(is^{t-1}) \sum_{a_j^{t-1}} Pr(a_j^{t-1}|\theta_j^{t-1}) O_i(s^t, a^{t-1}, o_i^t) \\ & T_i(s^{t-1}, a^{t-1}, s^t) \tau_{\theta_j^t}(b_j^{t-1}, a_j^{t-1}, o_j^t, b_j^t) O_j(s^t, a^{t-1}, o_j^t). \end{aligned} \tag{14}$$

We leave the proof of Propositions 3 to Section 4.

The value function $U(\theta_i)$ is

$$U(\theta_i, \mu_j) = \max_{a_i\in A_i} \left\{ \sum_{is} b_i(is) ER_i(is, a_i) + \gamma \sum_{o_i\in\Omega_i} Pr(o_i|a_i, b_i) U(\langle SE_{\theta_i}(b_i, a_i, o_i, \mu_j), \hat{\theta}_i) \right\} \tag{15}$$

where $ER_i(is, a_i) = \sum_{a_j} R_i(is, a_i, a_j) Pr(a_j|m_j, \mu_j)$.

And the set of optimal actions for agent i is defined as,

$$OPT(\theta_i, \mu_j) = \arg\max_{a_i\in A_i} \left\{ \sum_{is} b_i(is) ER_i(is, a_i) + \gamma \sum_{o_i\in\Omega_i} Pr(o_i|a_i, b_i) U(\langle SE_{\theta_i}(b_i, a_i, o_i, \mu_j), \hat{\theta}_i) \right\}. \tag{16}$$

Note the Equation (15) can be rewritten in the following form $U^n = HU^{n-1}$. Here $H : B \to B$ is a backup operator, and is defined as,

$$HU^{n-1}(\theta_i, \mu_j) = \max_{a_i\in A_i} h(\theta_i, a_i, \mu_j, U^{n-1}), \tag{17}$$

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

$$h(\theta_i, a_i, \mu_j, U) = \sum_{is} b_i(is) ER_i(is, a_i) + \gamma \sum_{o_i\in\Omega_i} Pr(o_i|a_i, b_i) U(\langle SE_{\theta_i}(b_i, a_i, o_i, \mu_j), \hat{\theta}_i). \tag{18}$$

3.1 ALGORITHM

Now we are ready to present the decentralized belief propagation algorithm for networked multi-agent systems. The algorithm requires each agent to maintain a belief on its interactive states, while allows each agent i share messages of certain type with its neighbors on the network. In this way, each agent is able to improve the value function and thus the current policy.

Decentralized Belief Propagation Algorithm

Input: Initialize $b_i(is)$, $U(\theta_i)$ for all $is \in IS_i$ and for all $i \in [N]:$ all i in $[N]$ Observe o_i^t, and reward r_i^t. Send message μ_i^t to all neighbors ∂i. Update the belief given the received messages $b_i^t = SE_{\theta_i}(b_i^{t-1}, a_i^{t-1}, o_i^t, \{\mu_j^t\}_{j\in\partial i})$. Update the value function $U(\theta_i) \leftarrow HU(\theta_i)$. Select and execute action a_i^t. convergence

Note similar to the typical belief propagation algorithm, all agents send/receive messages simultaneously and then update their beliefs simultaneously.

4 THEORETICAL RESULTS

We start this section by proving Proposition 3 for the case that message type is 'action', i.e., $M = \times_{i\in[N]} A_i$ and more specifically $\mu_j^t = a_j^{t-1}$. The belief update step for other message types can be derived in similar ways.

Proof 1 *We start by applying the Bayes Theorem:*

$$
\begin{aligned}
b_i^t(is^t) &= Pr(is^t|o_i^t, a_i^{t-1}, b_i^{t-1}, \mu_j^t) \\
&= \frac{Pr(is^t, o_i^t|a_i^{t-1}, b_i^{t-1}, \mu_j^t)}{Pr(o_i^t|a_i^{t-1}, b_i^{t-1}, \mu_j^t)} \\
&= \beta \sum_{is^{t-1}} b_i^{t-1}(is^{t-1}) Pr(is^t, o_i^t|a_i^{t-1}, is^{t-1}, \mu_j^t) \\
&= \beta \sum_{is^{t-1}} b_i^{t-1}(is^{t-1}) \sum_{a_j^{t-1}} Pr(a_j^{t-1}|a_i^{t-1}, is^{t-1}, \mu_j^t) Pr(is^t, o_i^t|a_i^{t-1}, a_j^{t-1}, is^{t-1}, \mu_j^t) \\
&= \beta \sum_{is^{t-1}} b_i^{t-1}(is^{t-1}) \sum_{a_j^{t-1}} Pr(a_j^{t-1}|is^{t-1}, \mu_j^t) Pr(is^t, o_i^t|a_i^{t-1}, a_j^{t-1}, is^{t-1}, \mu_j^t) \\
&= \beta \sum_{is^{t-1}} b_i^{t-1}(is^{t-1}) \sum_{a_j^{t-1}} Pr(a_j^{t-1}|b_j^{t-1}, \mu_j^t) Pr(o_i^t|a^{t-1}, is^t, is^{t-1}, \mu_j^t) Pr(is^t|a^{t-1}, is^{t-1}, \mu_j^t) \\
&\overset{(a)}{=} \beta \sum_{is^{t-1}} b_i^{t-1}(is^{t-1}) Pr(o_i^t|a^{t-1}, is^t) Pr(is^t|a^{t-1}, is^{t-1}) \\
&= \beta \sum_{is^{t-1}} b_i^{t-1}(is^{t-1}) O_i(s^t, a^{t-1}, o_i^t) Pr(is^t|a^{t-1}, is^{t-1}),
\end{aligned} \tag{19}
$$

where the equality (a) holds because $Pr(a_j^{t-1}|b_j^{t-1}, \mu_j^t) = 1$ when $\mu_j^t = a_j^{t-1}$ and $Pr(a_j^{t-1}|b_j^{t-1}, \mu_j^t) = 0$ otherwise. And recall $\mu_j^t = a_j^{t-1}$ is our assumption.

Since we assume the interactive states are intentional, $is^t = (s^t, \theta_j^t) = (s^t, b_j^t, \hat{\theta}_j^t)$, we can simplify the term $Pr(is^t|a^{t-1}, is^{t-1})$.

$$
\begin{aligned}
& Pr(is^t|a^{t-1}, is^{t-1}) \\
=& Pr(s^t, b_j^t, \hat{\theta}_j^t|a^{t-1}, is^{t-1}) \\
=& Pr(b_j^t|s^t, \hat{\theta}_j^t, a^{t-1}, is^{t-1}) Pr(s^t, \hat{\theta}_j^t|a^{t-1}, is^{t-1}) \\
=& Pr(b_j^t|s^t, \hat{\theta}_j^t, a^{t-1}, is^{t-1}) Pr(\hat{\theta}_j^t|s^t, a^{t-1}, is^{t-1}) Pr(s^t|a^{t-1}, is^{t-1}) \\
=& Pr(b_j^t|s^t, \hat{\theta}_j^t, a^{t-1}, is^{t-1}) I(\hat{\theta}_j^t, \hat{\theta}_j^{t-1}) T_i(s^{t-1}, a^{t-1}, s^t),
\end{aligned} \tag{20}
$$

where $I(\cdot,\cdot)$ is a boolean identity function, which equal 1 if the two frames are identical, and 0 otherwise. The joint action pair, a^{t-1}, may change the physical state. The third term on the right-hand side of Equation (20) captures this transition.

$$\begin{aligned}
&Pr(b_j^t|s^t,\hat{\theta}_j^t,a^{t-1},is^{t-1})\\
=&\sum_{o_j^t} Pr(b_j^t|s^t,\hat{\theta}_j^t,a^{t-1},is^{t-1},o_j^t)Pr(o_j^t|s^t,\hat{\theta}_j^t,a^{t-1},is^{t-1})\\
=&\sum_{o_j^t} Pr(b_j^t|s^t,\hat{\theta}_j^t,a^{t-1},is^{t-1},o_j^t)Pr(o_j^t|s^t,\hat{\theta}_j^t,a^{t-1})\\
=&\sum_{o_j^t} \tau_{\theta_j^t}(b_j^{t-1},a_j^{t-1},o_j^t,b_j^t)O_j(s^t,a^{t-1},o_j^t). \qquad (21)
\end{aligned}$$

In Equation (21), the first term on the right-hand side is 1 if agent j's belief update, $SE_{\theta_j}(b_j^{t-1},a_j^{t-1},o_j^t)$ generates a belief state equal to b_j^t. In the second terms on the right-hand side of the equation, the MNO assumption allows us to replace $Pr(o_j^t|s^t,\hat{\theta}_j^t,a^{t-1})$ with $O_j(s^t,a^{t-1},o_j^t)$.

Let us substitute Equation (21) into Equation (20),

$$Pr(is^t|a^{t-1},is^{t-1}) = \sum_{o_j^t} \tau_{\theta_j^t}(b_j^{t-1},a_j^{t-1},o_j^t,b_j^t)O_j(s^t,a^{t-1},o_j^t)I(\hat{\theta}_j^t,\hat{\theta}_j^{t-1})T_i(s^{t-1},a^{t-1},s^t). \qquad (22)$$

Now substitute Equation (22) into Equation (19), we have

$$\begin{aligned}
b_i^t(is^t) =&\beta \sum_{is^{t-1}} b_i^{t-1}(is^{t-1})\, O_i(s^t,a^{t-1},o_i^t) \sum_{o_j^t} \tau_{\theta_j^t}(b_j^{t-1},a_j^{t-1},o_j^t,b_j^t)O_j(s^t,a^{t-1},o_j^t)\\
&I(\hat{\theta}_j^t,\hat{\theta}_j^{t-1})T_i(s^{t-1},a^{t-1},s^t). \qquad (23)
\end{aligned}$$

We can remove the term $I(\hat{\theta}_j^t,\hat{\theta}_j^{t-1})$ by changing the scope of the first summation, which gives us the final expression for the belief update,

$$\begin{aligned}
b_i^t(is^t) =&\beta \sum_{is^{t-1}:\hat{m}_j^{t-1}=\hat{\theta}_j^t} b_i^{t-1}(is^{t-1})O_i(s^t,a^{t-1},o_i^t)T_i(s^{t-1},a^{t-1},s^t)\\
&\sum_{o_j^t} \tau_{\theta_j^t}(b_j^{t-1},a_j^{t-1},o_j^t,b_j^t)O_j(s^t,a^{t-1},o_j^t). \qquad (24)
\end{aligned}$$

Next, for an agent i and its I-POMDP$_i$, following the proof idea in (GD05), we prove the convergence of our algorithm. First, we show some properties of the back up operator H,

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

Proof 2 *Select arbitrary value functions V and U such that $V(\theta_i,\mu_V) \leq U(\theta_i,\mu_U)$, $\forall\theta_i \in \Theta_i$, $\mu_V,\mu_U \in M$, where θ_i is an arbitrary type of agent i and μ_V,μ_U are arbitrary messages.*

$$\begin{aligned}
&HV(\theta_i,\mu_V)\\
=&\max_{a_i\in A_i}\left\{\sum_{is} b_i(is)ER_i(is,a_i) + \gamma\sum_{o_i\in\Omega_i} Pr(o_i|a_i,b_i)V(\langle SE_{\theta_i}(b_i,a_i,o_i,\mu_V),\hat{\theta}_i)\right\}\\
=&\sum_{is} b_i(is)ER_i(is,a_i^*) + \gamma\sum_{o_i\in\Omega_i} Pr(o_i|a_i^*,b_i)V(\langle SE_{\theta_i}(b_i,a_i^*,o_i,\mu_V),\hat{\theta}_i)\\
\leq&\sum_{is} b_i(is)ER_i(is,a_i^*) + \gamma\sum_{o_i\in\Omega_i} Pr(o_i|a_i^*,b_i)U(\langle SE_{\theta_i}(b_i,a_i^*,o_i,\mu_U),\hat{\theta}_i)\\
\leq&\max_{a_i\in A_i}\left\{\sum_{is} b_i(is)ER_i(is,a_i) + \gamma\sum_{o_i\in\Omega_i} Pr(o_i|a_i,b_i)U(\langle SE_{\theta_i}(b_i,a_i,o_i,\mu_U),\hat{\theta}_i)\right\}\\
=&HU(\theta_i,\mu_U). \qquad (25)
\end{aligned}$$

Since θ_i,μ_V,μ_U are arbitrary, $HU \leq HV$.

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

Proof 3 *Assume two arbitrary well defined value functions V and U such that $V \leq U$. From Lemma 1, it follows that $HV \leq HU$. Let θ_i be an arbitrary type of agent i and μ_V, μ_U be arbitrary messages. And let a_i^* be the optimal action of $HU(\theta_i, \mu_U)$, we have,*

$$
\begin{aligned}
0 &\leq HV(\theta_i, \mu_V) - HU(\theta_i, \mu_U) \\
&= \max_{a_i \in A_i} \left\{ \sum_{is} b_i(is) ER_i(is, a_i) + \gamma \sum_{o_i \in \Omega_i} Pr(o_i | a_i, b_i) V(\langle SE_{\theta_i}(b_i, a_i, o_i, \mu_V), \hat{\theta}_i) \right\} \\
&- \max_{a_i \in A_i} \left\{ \sum_{is} b_i(is) ER_i(is, a_i) + \gamma \sum_{o_i \in \Omega_i} Pr(o_i | a_i, b_i) U(\langle SE_{\theta_i}(b_i, a_i, o_i, \mu_U), \hat{\theta}_i) \right\} \\
&\leq \sum_{is} b_i(is) ER_i(is, a_i^*) + \gamma \sum_{o_i \in \Omega_i} Pr(o_i | a_i^*, b_i) V(\langle SE_{\theta_i}(b_i, a_i^*, o_i, \mu_V), \hat{\theta}_i) \\
&- \sum_{is} b_i(is) ER_i(is, a_i^*) - \gamma \sum_{o_i \in \Omega_i} Pr(o_i | a_i^*, b_i) U(\langle SE_{\theta_i}(b_i, a_i^*, o_i, \mu_U), \hat{\theta}_i) \\
&= \gamma \sum_{o_i \in \Omega_i} Pr(o_i | a_i^*, b_i) \Big\{ V(\langle SE_{\theta_i}(b_i, a_i^*, o_i, \mu_V), \hat{\theta}_i) - U(\langle SE_{\theta_i}(b_i, a_i^*, o_i, \mu_U), \hat{\theta}_i) \Big\} \\
&= \gamma \sum_{o_i \in \Omega_i} Pr(o_i | a_i^*, b_i) \|V - U\| \\
&= \gamma \|V - U\|. \qquad (26)
\end{aligned}
$$

As the supremum norm is symmetrical, a similar result can be derived for $HU(\theta_i, \mu_U) - HV(\theta_i, \mu_V)$. Since θ_i, μ_V, μ_U are arbitrary, we prove the lemma.

Based on Lemma 1 and Lemma 2, following the Contraction Mapping Theorem in (Sto89), we can prove for each agent i, the value iteration in its I-POMDP$_i$ converges to a unique fixed point. We state the Contraction Mapping Theorem (Sto89) below

Theorem 1 (Contraction Mapping Theorem, (Sto89)) *If (S, ρ) is a complete metric space and $T : S \to S$ is a contraction mapping with modulus γ, then*

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

Theorem 2 *For a networked multi-agent I-POMDP, Algorithm 3.1 converges if the value functions of all agents are well defined.*

Proof 4 *First, the normed space $(B, \|\cdot\|)$ is complete w.r.t. the metric induced by the supremum norm. Second, Lemma 2 proves the contraction property of the operator H. Directly applying Theorem 1, letting $T = H$, we prove the value iteration in I-POMDPs converges to a unique fixed point.*

And we naturally have the following theorem.

Theorem 3 *For a networked multi-agent I-POMDP, the optimal policies for agent i, $i \in [N]$ is given by Equation (16).*

5 APPLICATIONS

In this section, we show our networked multi-agent I-POMDPs framework can be applied to various applications.

5.1 DECENTRALIZED CONTROL PROBLEM

Let us consider the partial history sharing information model in (NMT13). Consider a dynamic system with N controllers. The system operates in discrete time for a horizon T. Let $X^{(t)} \in \mathcal{X}^{(t)}$ denote the state of the system at time t, $U_i^{(t)} \in \mathcal{U}_i^{(t)}$ denote the control action of controller i, $i \in [N]$ at time t, and $\mathbf{U}^{(t)}$ denote the vector $(U_1^{(t)}, \dots, U_N^{(t)})$. The initial state $X^{(1)}$ has a probability distribution $Q^{(1)}$ and evolves according to

$$X^{(t+1)} = f^{(t)}(X^{(t)}, \mathbf{U}^{(t)}, W_0^{(t)}), \tag{27}$$

where $\{W_0^{(t)}\}_{t=1}^T$ is a sequence of i.i.d. random variables with probability distribution $Q_{W,0}$.

At any time t, each controller has access to three types of data: current observation, local memory, and shared memory.

- Current local observation: Each controller makes a local observation $Y_i^{(t)} \in \mathcal{Y}_i^{(t)}$ on the state of the system at time t,

$$Y_i^{(t)} = h_i^{(t)}(X^{(t)}, W_i^{(t)}), \tag{28}$$

  where $\{W_i^{(t)}\}_{t=1}^T$ is a sequence of i.i.d. random variables with probability distribution $Q_{W,i}$. We assume that the random variables in the collection $\{X^{(1)}, W_j^{(t)}, t = 1, \dots, T, j = 0, 1, \dots, N\}$ are mutually independent.

- Local memory: Each controller stores a subset $M_i^{(t)}$ of its past local observations and its past actions in a local memory:

$$M_i^{(t)} \subset \{Y_i^{(1:t)}, U_i^{(1:t)}\}. \tag{29}$$

  At $t = 1$, the local memory is empty, $M_1^{(t)} = \emptyset$.

- Shared memory: In addition to its local memory, each controller has access to a shared memory. The contents C_t of the shared memory at time t are a subset of the past local observations and control actions of all controllers:

$$C^{(t)} \subset \{\mathbf{Y}^{(1:t)}, \mathbf{U}_i^{(1:t)}\}, \tag{30}$$

  where $\mathbf{Y}^{(t)}$ and $\mathbf{U}^{(t)}$ denote the vectors $(Y_1^{(t)}, \dots, Y_N^{(t)})$ and $(U_1^{(t)}, \dots, U_N^{(t)})$ respectively. At $t = 1$, the shared memory is empty, $C^{(1)} = \emptyset$.

Controller i chooses action $U_i^{(t)}$ as a function of the total data $Y_i^{(t)}, M_i^{(t)}, C^{(t)}$ available to it. Specifically, for every controller i, $i \in [N]$,

$$U_i^{(t)} = g_i^{(t)}(Y_i^{(t)}, M_i^{(t)}, C^{(t)}), \tag{31}$$

where $g_i^{(t)}$ is called the control law of controller i. The collection $\mathbf{g}_i = (g_i^{(1)}, \dots, g_i^{(T)})$ is called the control strategy of controller i. The collection $\mathbf{g}_{1:N} = (\mathbf{g}_1, \dots, \mathbf{g}_N)$ is called the control strategy of the system.

At time t, the system incurs a cost $l(X^{(t)}, \mathbf{U}^{(t)})$. The performance of the control strategy of the system is measured by the expected total cost

$$J(\mathbf{g}_{1:N}) := \mathbb{E}^{\mathbf{g}_{1:N}}[\sum_{t=1}^{T} l(X^{(t)}, \mathbf{U}^{(t)})], \tag{32}$$

where the expectation is with respect to the joint probability measure on $(X^{(1:T)}, \mathbf{U}^{(1:T)})$ induced by the choice of $\mathbf{g}_{1:N}$.

We are interested in the following optimization problem

Definition 6 *For the model described above, given the state evolution functions $f^{(t)}$, the observation functions $h_i^{(t)}$, the protocols for updating local and share memory, the cost function l, the distributions $Q^{(1)}$, $Q_{W,i}$, $i = 0, 1, \dots, N$, and the horizon T, find a control strategy $\mathbf{g}_{1:N}$ for the system that minimized the expected total cost given by Equation (32).*

(NMT13) show the decentralized system defined in Definition 6 can be viewed as a coordinated system. The coordinator only knows the shared memory $C^{(t)}$ at time t. At time t, the coordinator chooses mappings $\Gamma_i^{(t)} : \mathcal{Y}_i^{(t)} \times \mathcal{M}_i^{(t)} \to \mathcal{U}_i^{(t)}$, for $i \in [n]$, according to

$$\mathbf{\Gamma}^{(t)} = d^{(t)}(C^{(t)}, \mathbf{\Gamma}^{(1:t\text{-}1)}), \tag{33}$$

where $\mathbf{\Gamma}^{(t)} = (\Gamma_1^{(t)}, \Gamma_2^{(t)}, \ldots, \Gamma_n^{(t)})$, and the function $d^{(t)}$ is called *coordination rule* at time t. The function $\Gamma_i^{(t)}$ is called the *coordinator's prescription* to controller i. At time t, the function $\Gamma_i^{(t)}$ is communicated to controller i, and then the controller i generates an action using the function $\Gamma_i^{(t)}$ based on its current local observation and its local memory:

$$U_i^{(t)} = \Gamma_i^{(t)}(Y_i^{(t)}, M_i^{(t)}). \tag{34}$$

Moreover, the coordinated system can further be viewed as an instance of a POMDP model by defining the state process as $s^t := \{X^{(t)}, \mathbf{Y}^{(t)}, \mathbf{M}^{(t)}\}$, the observation process as $o^t := C^{(t\text{-}1)}$, and the action process $A^t := \mathbf{\Gamma}^{(t)}$. And we can define the information state at time t for the POMDP of the coordinator as:

$$\Pi^{(t)} := \mathbb{P}(s^{(t)} | C^{(t)}, \mathbf{\Gamma}^{(1:t)}). \tag{35}$$

Furthermore, we have a new system dynamic at time t as

$$\Pi^{(t+1)} = \eta^{(t)}(\Pi^{(t)}, C^{(t)}, \mathbf{\Gamma}^{(t)}), \tag{36}$$

where $\eta^{(t)}$ is the standard non-linear filtering update function (see (NMT13) for more details).

Now, given our framework in Section 3, we can prove that the above optimization problem is a networked multi-agent I-POMDP, as shown in the following proposition. The optimization problem defined in Definition 6 is a networked multi-agent I-POMDP.

Proof 5 *We can prove the proposition by defining the networked multi-agent I-POMDP tuple* $\langle \mathcal{G}, \{IS_i\}_{i\in[N]}, \{A_i\}_{i\in[N]}, \{T_i\}_{i\in[N]}, \{\Omega_i\}_{i\in[N]}, \{O_i\}_{i\in[N]}, \{R_i\}_{i\in[N]}, M \rangle$ *for the optimization problem,*

- *Given the definition of shared memory, we can define an equivalent communication network* $\mathcal{G} = (\mathcal{V}, \mathcal{E})$*. Let the set of* N *agents (controllers) be the set of nodes* $\mathcal{V}$*, labeled by index* $i = 1, 2, ..., N$*. And an edge* (i, j) *is in* $\mathcal{E}$ *if and only if agent* i *shares memory with agent* j*.*
- *For each agent (controller)* i*, the interactive state* $IS_i = \mathcal{X} \times M_{\partial i}$*, where* $\mathcal{X}$ *is the set of states of the physical environment, and* $M_{\partial i}$ *is the set of possible models of* i*'s neighbors* ∂i*.*
- *For each agent (controller)* i*, the action space* $A_i^{(t)}$ *at time* t *is given by the range of the coordinator's prescription to controller* i*,* $\Gamma_i^{(t)}$*.*
- *For each agent (controller)* i*, the transition model* $T_i^{(t)}$ *at time* t *is given by the dynamic* $\Pi^{(t+1)} = \eta^{(t)}(\Pi^{(t)}, C^{(t)}, \mathbf{\Gamma}^{(1:t\text{-}1)})$*.*
- *For each agent (controller)* i*, the set of observations* $\Omega_i^{(t)}$ *at time* t *is given by the set of possible shared memories* $\{C^{(t)}\}$*.*
- *For each agent (controller)* i*, the observation function* $O_i^{(t)}$ *at time* t *is given by* $Y_i^{(t)} = h_i^{(t)}(X^{(t)}, W_i^{(t)})$ *and* $\Gamma_i^{(t)}$*.*
- *For each agent (controller)* i*, the reward function at time* t *is given by* $l(X^{(t)}, \mathbf{U}^{(t)})$*.*
- *The message type depends on the definition of shared memory, it could be 'action' or 'observation'.*

The optimal strategy of the optimization problem in Definition 6 can be obtained by running Algorithm 3.1 and given by Theorem 3.

(NMT13) call their solution Dynamic Programming Decomposition. We generalize the scenario to the networked multi-agent case, and may call our solution Belief Propagation Decomposition.

5.2 DECENTRALIZED SPECTRUM SHARING PROBLEM

We consider a contention based decentralized spectrum sharing problem. Given a communication network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, such that $\mathcal{V}$ is a set of N base stations, labeled by an index $i = 1, 2, \ldots, N$, and an edge (i, j) is in $\mathcal{E}$ if and only if base station i is backhauled with base station j. Each base station serves a given subset of user equipments (UEs). The whole communication network shares a single spectrum, the base stations contend the transmission opportunities in the following way. Each time slot consists of two phases, contention phase and data transmission phase. At contention phase, each base station draws a random number at the start of a time slot, which determines the order of optional transmissions. In the designated slot of the contention phase, a base station can choose to transmit or keep silent. If a base station transmits, it continues transmission through the contention phase and data transmission phase. Ideally, the UE throughput is given by Shannon channel capacity. And the objective of each base station is to maximize the long-term throughput it delivers to its UEs.

For simplicity, we make the following assumptions. We assume each base station serves only one UE. Each base station always has traffic to be delivered to the UE, thus always participates in contention. And there is only downlink traffic. The action space of each base station is $\{\text{transmit}, \text{silent}\}$, which can be denoted as $\{1, 0\}$.

Given the above assumptions, we can mathematically formulate the problem. Let us denote the UE served by the base station i by the same index i, and so is the link between the UE and the base station. We denote the link strength between base station i and UE j by channel coefficient h_{ij}. For each link i in time slot t, let us denote the transmission rate by $R_i^{(t)}$ and the long-term average rate by $\bar{X}_i^{(t)}$.

$$R_i^{(t)} = W \log_2(1 + \text{SINR}_i^{(t)}), \tag{37}$$

$$\bar{X}_i^{(t)} = (1 - \frac{1}{B})\bar{X}_i^{(t\text{-}1)} + \frac{1}{B} R_i^{(t)}, \tag{38}$$

where $B > 1$ is a parameter which balances the weights of past and current transmission rates. We denote the actions of all base stations in time slot t as $\mathbf{a}^{(t)} = [a_1^{(t)}, \ldots, a_N^{(t)}]^{\in} \{0, 1\}^N$. The SINR for UE i is given by

$$\text{SINR}_i^{(t)} = \frac{h_{ii}^{(t)} P_t a_i^{(t)}}{\sigma_{\text{UE}}^2 + \sum_{j \neq i} h_{ji}^{(t)} P_t a_j^{(t)}} = \frac{S_i^{(t)}}{\sigma_{\text{UE}}^2 + I_i^{(t)}}, \tag{39}$$

where P_t is the transmission power, and σ_{UE}^2 is the noise power at UE. $S_i^{(t)}$ is the signal power for UE i at time t, and $I_i^{(t)}$ is the total interference power for UE i at time t.

Given the action vector $\mathbf{a}^{(t)}$ in each time slot t, the long term proportional fairness scheduling utility is (KMT98)

$$\max_{t \to \infty} U(\bar{\mathbf{X}}^{(t)}) = \max_{t \to \infty} \sum_{i=1}^{N} \log(\bar{X}_i^{(t)}). \tag{40}$$

And we can split the proportional fairness metric over time by rewriting the utility function up to time slot T,

$$\begin{aligned} U(\bar{\mathbf{X}}^{(T)}) &= \sum_{i=1}^{N} \log(\bar{X}_i^{(T)}) \\ &= \sum_{i=1}^{N} \log\left((1-\frac{1}{B})\bar{X}_i^{(T\text{-}1)} + \frac{1}{B}R_i^{(T)}\right) \\ &= \sum_{i=1}^{N} \log\left((1-\frac{1}{B})\bar{X}_i^{(T\text{-}1)}(1+\frac{R_i^{(T)}}{(B-1)\bar{X}_i^{(T\text{-}1)}})\right) \\ &= \sum_{i=1}^{N} (\log \bar{X}_i^{(0)} + \sum_{t=1}^{T} r_i^{(t)}) \\ &= \sum_{i=1}^{N} \log \bar{X}_i^{(0)} + \sum_{t=1}^{T}\sum_{i=1}^{N} r_i^{(t)}, \end{aligned} \tag{41}$$

where

$$r_i^{(t)} = \log\left((1-\frac{1}{B})\left(1+\frac{R_i^{(T)}}{(B-1)\bar{X}_i^{(T\text{-}1)}}\right)\right). \tag{42}$$

Given our framework, we can prove that the above optimization problem is a networked multi-agent I-POMDP, as shown in the following proposition.

The decentralized spectrum sharing problem defined in Section 5.2 is a networked multi-agent I-POMDP.

Proof 6 *We can prove the proposition by defining the networked multi-agent I-POMDP tuple* $\langle \mathcal{G}, \{IS_i\}_{i\in[N]}, \{A_i\}_{i\in[N]}, \{T_i\}_{i\in[N]}, \{\Omega_i\}_{i\in[N]}, \{O_i\}_{i\in[N]}, \{R_i\}_{i\in[N]}, M\rangle$ *for the optimization problem,*

- *The communication network* $\mathcal{G}$ *is directly defined in the decentralized spectrum sharing problem.*
- *For each agent (base station)* i*, the interactive state* $IS_i = S \times M_{\partial i}$*, where* S *consists of the joint space of the average rate of link* i *and the channels between all base stations and UE* i*, i.e.* $\langle \bar{X}_i^{(t)}, \{h_{ji}\}_{j\in[N]}\rangle$ *is a state in* S*, and* $M_{\partial i}$ *is the set of possible models of* i*'s neighbors* ∂i*.*
- *For each agent (base station)* i*, the action space* A_i *is* {*transmit, silent*}*, i.e.,* $\{0, 1\}$*.*
- *For each agent (base station)* i*, the transition model* $T_i : S \times A_i \times A_{\partial i} \times S \to [0, 1]$ *is given by the dynamic*

  $$\bar{X}_i^{(t)} = (1-\frac{1}{B})\bar{X}_i^{(t\text{-}1)} + \frac{1}{B}R_i^{(t)},$$

  and the channel fading model.
- *For each agent (base station)* i*, at time slot* t*, the observation consists of the average throughput* $\bar{X}_i^{(t\text{-}1)}$*, the signal power* $S_i^{(t\text{-}1)}$ *and the total interference power* $I_i^{(t\text{-}1)}$ *of the previous time slot, i.e.,* $o_i^t = [\bar{X}_i^{(t\text{-}1)}, S_i^{(t\text{-}1)}, I_i^{(t\text{-}1)}]$*.*
- *For each agent (base station)* i*, the observation function* O_i *is directly given by the definitions of involving parameters in* S*,* $A_i \times A_{\partial i}$ *and* Ω_i
- *For each agent (controller)* i*, at time slot* t*, the reward function is given by* $r_i^{(t)}$*.*
- *The message type in this problem can be either 'action' or 'observation'.*

The Proposition 5.2 naturally leads to the following corollary, The optimal strategy of decentralized spectrum sharing problem can be obtained by running Algorithm 3.1 and given by Theorem 3.

6 CONCLUSION

In this paper, we address the problem of multi-agent I-POMDPs with networked agents. In particular, we consider the fully decentralized setting where each agent makes individual decisions and receives local rewards, while exchanging information with neighbors over the network to accomplish optimal network-wide averaged return. Within this setting, we propose a decentralized belief propagation algorithm. We provide theoretical analysis on the convergence of the proposed algorithm. And we show our framework can be applied to various applications. An interesting direction of future research is to extend our algorithms and analyses to the policy gradient methods.

REFERENCES

[AB02] Jeffrey L Adler and Victor J Blue. A cooperative multi-agent transportation management and route guidance system. *Transportation Research Part C: Emerging Technologies*, 10(5-6):433–454, 2002.

[AO15] Christopher Amato and Frans A Oliehoek. Scalable planning and learning for multiagent pomdps. Technical report, MASSACHUSETTS INST OF TECH CAMBRIDGE COMPUTER SCIENCE AND ARTIFICIAL ..., 2015.

[ASSC02] Ian F Akyildiz, Weilian Su, Yogesh Sankarasubramaniam, and Erdal Cayirci. Wireless sensor networks: a survey. *Computer networks*, 38(4):393–422, 2002.

[BBX+16] Dzmitry Bahdanau, Philemon Brakel, Kelvin Xu, Anirudh Goyal, Ryan Lowe, Joelle Pineau, Aaron Courville, and Yoshua Bengio. An actor-critic algorithm for sequence prediction. *arXiv preprint arXiv:1607.07086*, 2016.

[BDH99] Craig Boutilier, Thomas Dean, and Steve Hanks. Decision-theoretic planning: Structural assumptions and computational leverage. *Journal of Artificial Intelligence Research*, 11:1–94, 1999.

[BGIZ02] 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.

[DVPWR13] Finale Doshi-Velez, David Pfau, Frank Wood, and Nicholas Roy. Bayesian nonparametric methods for partially-observable reinforcement learning. *IEEE transactions on pattern analysis and machine intelligence*, 37(2):394–407, 2013.

[FADFW16] Jakob Foerster, Ioannis Alexandros Assael, Nando De Freitas, and Shimon Whiteson. Learning to communicate with deep multi-agent reinforcement learning. In *Advances in neural information processing systems*, pages 2137–2145, 2016.

[FNF+17] Jakob Foerster, Nantas Nardelli, Gregory Farquhar, Triantafyllos Afouras, Philip HS Torr, Pushmeet Kohli, and Shimon Whiteson. Stabilising experience replay for deep multi-agent reinforcement learning. *arXiv preprint arXiv:1702.08887*, 2017.

[GD05] Piotr J Gmytrasiewicz and Prashant Doshi. A framework for sequential planning in multi-agent settings. *Journal of Artificial Intelligence Research*, 24:49–79, 2005.

[GEK17] Jayesh K Gupta, Maxim Egorov, and Mykel Kochenderfer. Cooperative multi-agent control using deep reinforcement learning. In *International Conference on Autonomous Agents and Multiagent Systems*, pages 66–83. Springer, 2017.

[HG18] Yanlin Han and Piotr Gmytrasiewicz. Learning others' intentional models in multi-agent settings using interactive pomdps. In *Advances in Neural Information Processing Systems*, pages 5634–5642, 2018.

[HW03] Junling Hu and Michael P Wellman. Nash q-learning for general-sum stochastic games. *Journal of machine learning research*, 4(Nov):1039–1069, 2003.

[KLC98] Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134, 1998.

[KMT98] Frank P Kelly, Aman K Maulloo, and David KH Tan. Rate control for communication networks: shadow prices, proportional fairness and stability. *Journal of the Operational Research society*, 49(3):237–252, 1998.

[Lit94] Michael L Littman. Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994*, pages 157–163. Elsevier, 1994.

[Lit01] Michael L Littman. Value-function reinforcement learning in markov games. *Cognitive systems research*, 2(1):55–66, 2001.

[LLC11] Miao Liu, Xuejun Liao, and Lawrence Carin. The infinite regionalized policy representation. In *ICML*, 2011.

[LR00] Martin Lauer and Martin Riedmiller. An algorithm for distributed reinforcement learning in cooperative multi-agent systems. In *In Proceedings of the Seventeenth International Conference on Machine Learning*. Citeseer, 2000.

[MSL11] Joao V Messias, Matthijs Spaan, and Pedro U Lima. Efficient offline communication policies for factored multiagent pomdps. In *Advances in Neural Information Processing Systems*, pages 1917–1925, 2011.

[NMT13] Ashutosh Nayyar, Aditya Mahajan, and Demosthenis Teneketzis. Decentralized stochastic control with partial history sharing: A common information approach. *IEEE Transactions on Automatic Control*, 58(7):1644–1658, 2013.

[NTY+03] Ranjit Nair, Milind Tambe, Makoto Yokoo, David Pynadath, and Stacy Marsella. Taming decentralized pomdps: Towards efficient policy computation for multiagent settings. In *IJCAI*, volume 3, pages 705–711, 2003.

[OPA+17] Shayegan Omidshafiei, Jason Pazis, Christopher Amato, Jonathan P How, and John Vian. Deep decentralized multi-task multi-agent reinforcement learning under partial observability. *arXiv preprint arXiv:1703.06182*, 2017.

[OW96] James M Ooi and Gregory W Wornell. Decentralized control of a multiple access broadcast channel: Performance bounds. In *Proceedings of 35th IEEE Conference on Decision and Control*, volume 1, pages 293–298. IEEE, 1996.

[PG16] Alessandro Panella and Piotr Gmytrasiewicz. Bayesian learning of other agents' finite controllers for interactive pomdps. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 2530–2536, 2016.

[PT87] Christos H Papadimitriou and John N Tsitsiklis. The complexity of markov decision processes. *Mathematics of operations research*, 12(3):441–450, 1987.

[PT02] David V Pynadath and Milind Tambe. The communicative multiagent team decision problem: Analyzing teamwork theories and models. *Journal of artificial intelligence research*, 16:389–423, 2002.

[RN02] Stuart Russell and Peter Norvig. Artificial intelligence: a modern approach. 2002.

[RN04] Michael Rabbat and Robert Nowak. Distributed optimization in sensor networks. In *Proceedings of the 3rd international symposium on Information processing in sensor networks*, pages 20–27, 2004.

[Sto89] Nancy L Stokey. *Recursive methods in economic dynamics*. Harvard University Press, 1989.

[ZYL+18] Kaiqing Zhang, Zhuoran Yang, Han Liu, Tong Zhang, and Tamer Başar. Fully decentralized multi-agent reinforcement learning with networked agents. *arXiv preprint arXiv:1802.08757*, 2018.