Asking for information by evaluating the communication and coordination trade-off in multi-agent POMDPs

***Abstract*— This letter concerns decentralized multi-agent systems in which individual agents operate under partial observability and uncertainty, but with the option of invoking inter-agent communication. When information exchange is infeasible or prohibitively expensive—owing to bandwidth limits, energy constraints, or the demands of clandestine or contested operation— one is prevented from continually maintaining a joint belief over the underlying system state. In such cases, the decision of when to request information from others is key. The currently existing strategies are conservative: agents are constrained to act solely on information that is accessible to all, with useful local observations being ignored whenever communication is too costly. Such an approach, though apposite to high-risk settings when consistent decision-making must be guaranteed, will oftentimes degrade overall performance significantly. We explore how agents may integrate their local information, even when it has not been globally shared, and then communicate only when prudent. We propose a novel method that enables the decentralized execution of a centralized policy by reasoning, at execution time, about whether the cost of exchanging information will be offset by the information's value. We introduce an *ask-type* paradigm in which agents are not compelled to synchronize, and we empirically evaluate our approach across several benchmarks, assessing in each case its effectiveness in trading between tolerating some loss of coordination and reducing communication costs.**

I. Introduction

Decision-making under uncertainty for multiple agents is a focus area of research that continues to grow owing to its increasing relevance to applications such as target tracking, surveillance, sensor networks, cooperative robotics, intelligent transportation, task allocation, collision-free navigation, and cooperative exploration, among others [1]. *Cooperative decision-making* refers to the challenge of deriving a set of local behaviors for a group of agents operating within an environment to maximize a global reward. In decentralized systems, individual decision-makers lack access to global information and must instead rely on their local histories of observations and actions while reasoning about the potential behaviors of other agents. This problem is formalized by the Decentralized Partially Observable Markov Decision Process (Dec-POMDP) framework [2], which generalizes the Partially Observable Markov Decision Process (POMDP) [3] to a multi-agent context. In practice, the strict decentralization assumption can oftentimes be relaxed, allowing agents to exchange information whenever doing so is beneficial. During mission execution, inter-agent communication serves as a coordination mechanism through which agents synchronize their beliefs, mitigate uncertainty, and improve performance, potentially approaching the effectiveness of a fully centralized planning while still maintaining decentralized execution. When communication is free and broadcasted, the problem reduces to a large centralized POMDP, known as a Multi-agent Partially Observable Markov Decision Process (MPOMDP), where the fully centralized policy is executed through continuous information exchange. Unfortunately, continuous communication can be infeasible or undesirable due to bandwidth limitations, energy constraints, or the risk of detection in hostile environments.

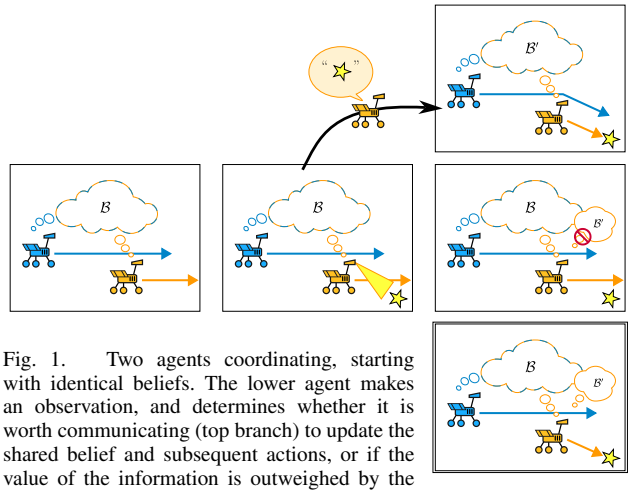


Fig. 1. Two agents coordinating, starting with identical beliefs. The lower agent makes an observation, and determines whether it is worth communicating (top branch) to update the shared belief and subsequent actions, or if the value of the information is outweighed by the cost of communication. In the latter case, the joint actions continue to be derived from the joint belief. We examine (double box at right) when the agent uses its local belief even though this may cause deviations from what is known collectively.

A. State of the art

Existing solution methods, both model-based and learning-based, seek to optimize action-selection and communication either during offline planning or online execution. In this paper, we investigate model-based approaches grounded in decision theory, and we focus on planning strategies that explicitly account for communication to (i) enable decentralized execution of a centralized policy, or (ii) enhance performance during decentralized policy execution.

Deciding *when* to communicate is crucial to avoid bottlenecks, wasted computational resources, and redundant information. Existing approaches include treating communication as an available action in the framework of Dec-POMDPs, which can be solved offline [4], [5] or online (e.g. [6]); agent-centric interactive POMDP formulations, where agents recursively reason about other agents' beliefs [7]–[9]; and game-theoretic approaches [10].

An important early piece of work, that of Roth, Simmons & Veloso [11], approaches the problem of decentralized coordination by deliberating, during execution, over whether it is beneficial to share some information with

other agents (they do assume that free communication is available at planning time). Agents maintain a joint belief that is consistent with information that has been shared with all. When an agent learns something new —e.g., disclosed via a recent local observation— it determines, depending on whether the communication cost can be justified, if it is worth everyone knowing it or not. When the answer is not to communicate, the agent with the new information acts as if itself were unaware (See Fig. 1). This approach, wherein the onus to be consistent with one's peers overrides one's own local knowledge, is *conservative*; it is an appropriate choice in high-risk circumstances. But can we obtain more efficient behavior by allowing agents to use local information? We show that there are task domains in which agents updating their behavior on the basis of local information, even at the potential risk of misalignment, can be beneficial so long as the agents reason about the impact of such misalignment. Indeed, several standard benchmark problems are seen to have this property.

B. Contribution and Organization

Our contributions are summarized as follows:

- We introduce a novel algorithm that enables decentralized execution of a centralized policy by reasoning, at each time step, about communication.
- We propose an ask-type communication paradigm in which each agent preserves an individual perspective during execution, integrates local information, and requests information from the team only when necessary. By adopting a *broadcast query strategy*, our method focuses on optimizing *when* to communicate, enabling agents to reduce uncertainty and improve local action selection while minimizing communication overhead.
- Finally, we evaluate the proposed algorithm on benchmarks from the literature, showing its effectiveness in managing the trade-off between coordination loss and communication cost.

The remainder of this paper is organized as follows: Section II presents the problem formulation; Section III describes the proposed approach; Section IV reports and discusses experimental results; and Section V concludes the paper.

II. Problem formulation

Throughout this work, we consider a decentralized multi-agent system composed of rational, cooperative agents operating in a stochastic, partially observable environment. Agents possess local information and may establish a *joint belief* about the global state of the system through communication, which is assumed to be always available, instantaneous, and reliable, but costly. In the following sections, we formally introduce the planning problem, describe belief space generation and propagation, and analyze the proposed strategy to solve the planning problem and mitigate its complexity.

A. System model

An MPOMDP (or, equivalently, a Dec-POMDP) is defined as a tuple:

$$\langle I, S, A, T, O, \mathcal{O}, R, b_0, h, \gamma \rangle,$$

where $I = \{1, \dots, n\}$ is the set of n agents, S is a finite set of states, summarizing all the relevant features of the agents and their environment, $A = \times_{i \in I} A_i$ is the finite set of joint actions and $O = \times_{i \in I} O_i$ the finite set of joint observations. (The $\times$ operator denotes a Cartesian product.) A joint action for the current time step t is denoted as $a_t = \langle a_{1,t}, \dots, a_{n,t} \rangle$. Similarly, $o_t = \langle o_{1,t}, \dots, o_{n,t} \rangle$ denotes a joint observation for the current time step t. Every time step, agent i only observes its own component o_i.[1] Let $\langle o_i, o_{-i} \rangle$ denote the set of joint observations with o_i the local observation and $o_{-i} \in O_{-i} = \times_{j \in I,\ j \neq i} O_j$ observations of the other agents except i.[2]

The transition probability function and the observation probability function are denoted as T and $\mathcal{O}$, respectively. In particular, $T(s'|s, a)$ denotes the probability that taking joint action a in state s results in a transition to state s'. $\mathcal{O}(o|s', a)$ denotes the probability of obtaining joint observation o after taking joint action a and reaching state s'. Transition and observation probabilities are assumed to be stationary.

The reward function $R(s, a)$ defines the reward received when the system executes joint action a in state s. The initial state distribution or *belief* at time $t = 0$ is denoted by $b_0 \in \Delta(S)$, with $\Delta(S)$ the set of probability distributions over the set of states. All agents share the initial belief and have complete knowledge of the system model, including the transition dynamics and observation functions.

Finally, h and $\gamma \in [0, 1)$ are the time horizon and the discount factor, respectively.

The distinctive feature of an MPOMDP is the ability of each agent to compute a *joint belief*,

$$b_t = \Pr(s_t | b_0, a_0, o_1, a_1, \dots, a_{t-1}, o_t),$$

defined as the probability distribution over states given the histories of joint actions and observations [2].

An action-observation history (AOH) for agent i at time step t, $\vec{\theta}_{i,t}$, is the sequence of actions taken by and observations received by agent i,

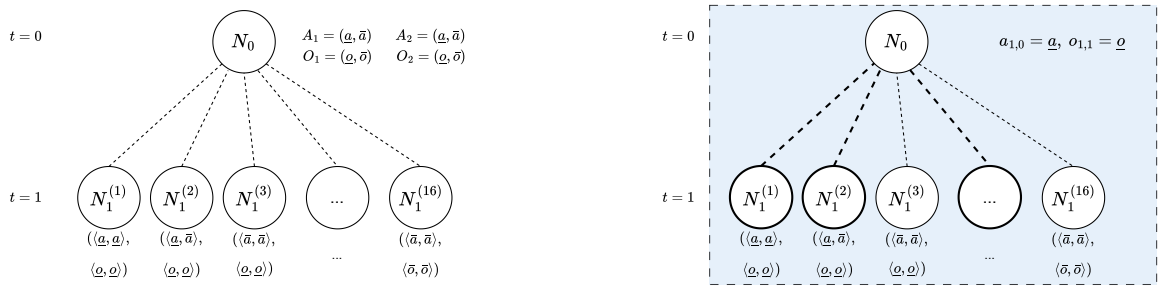
$$\vec{\theta}_{i,t} = (a_{i,0}, o_{i,1}, \dots, a_{i,t-1}, o_{i,t}).$$

Agent i's set of possible AOHs at time t is $\vec{\Theta}_{i,t}$. A joint action-observation history $\vec{\theta}_t$ specifies the AOH for all agents at time t, while the set of possible joint AOHs for all agents is denoted as $\vec{\Theta}_t$. Similar definitions are used when defining the agent i and the joint observation history (OH) for the sequence of observations and joint observations received by the agent i and the team, respectively.

The joint belief b_t serves as a *sufficient statistic* for the system, enabling agents to select actions based on the current state of knowledge. Its computation can be done

[1] For conciseness, the time index is omitted.

[2] With the notation $-i$, we denote the set of all agents except agent i.



(a) System overview

(b) Local overview, Agent 1

Fig. 2. A common initial root N_0 represents the initial state of a system composed of two agents with local action and observation spaces $A_i = (\underline{a}, \bar{a})$, and $O_i = (\underline{o}, \bar{o})$, respectively, $i = 1, 2$. A branching structure is generated to account for all possible action-observation combinations (Fig. 2a). Each agent can isolate the branches consistent with their own local AOH (Fig. 2b).

incrementally using Bayes' rule in the same way as in a POMDP, i.e., through a Bayesian belief update function $\tau(b_{t-1}, a_{t-1}, o_t)$ that computes b_t from b_{t-1}, a_{t-1}, and o_t.

Solving an MPOMDP involves defining a plan, or *policy*, for each agent that maximizes the expected cumulative reward:[3]

$$E\Big[\sum_{t=0}^{h-1} R(s_t, a_t) | b_0, \pi\Big],$$

also referred to as the *value* V^π of a joint policy π, where the expectation is over states and observations.

The policy $\pi = (\delta_0, \ldots, \delta_{h-1})$ maps joint beliefs to joint actions at each time step t,

$$\delta_t : \Delta(S) \to A \ \forall t,$$

and can be decomposed into individual policies $\pi_i = (\delta_{i,0}, \ldots, \delta_{i,h-1})$, with $\delta_{i,t} : \Delta(S) \to A_i \ \forall t$ for all agents. In §II-C, we discuss methods for identifying the optimal policy, highlighting the dual challenge of optimizing policy computation and execution whilst simultaneously minimizing costly communication.

B. Belief space generation and propagation

To identify the optimal action, each agent must reconstruct a joint belief over the system state by exchanging local information with other agents at every time step. If this exchange is not possible, agents cannot fully reconstruct their teammates' AOHs and, as a result, cannot determine a unique joint belief. They can only identify possible joint AOHs —and therefore possible joint beliefs— which can be visualized as branches of a tree starting from a common root (Fig. 2), where each path represents a specific joint history, its resulting joint belief, and its likelihood of occurrence.

For each agent i, let $\mathcal{B}_t^i$ denote the set of possible joint beliefs compatible with its local information. This set is constructed by agent i based on its own AOH and on reasoning about the possible AOHs of the other agents. At time $t = 0$, all $\mathcal{B}_0^i$ are identical and consist solely of the initial belief b_0. From the joint belief b_0, agent i can derive an optimal joint action a_0 from the centralized policy π and execute its corresponding individual action $a_{i,0}$. The system then transitions to a new state, and each agent receives a local observation $o_{i,1}$.

If the agents decide to communicate (Fig. 3a), they can reconstruct the joint AOH and update the joint belief accordingly. All sets $\mathcal{B}_1^i$ are restricted to the current joint belief $b_1 = \tau(b_0, a_0, o_1)$ about the current joint AOH $\vec{\theta}_1 = (a_0, o_1)$, whose probability of occurrence is $\Pr(\vec{\theta}_1) = 1$. If agents do not communicate (Fig. 3b), the exact joint belief cannot be constructed, but a set of possible joint beliefs can be inferred. Agent i might naturally do one of two things. First, (i) act blindly and enumerate all possible joint OHs from the root and all possible joint AOHs otherwise–the root constitutes a special case as a_0 is coordinated. Or (ii) incorporate its local information and infer the possible OHs of the remaining agents—in subsequent steps, their AOHs. The first option ensures that all agents reason from the same information–coordination is guaranteed, but valuable local information can be wasted. The second option leverages local information to improve decision-making—even at the potential cost of temporary misalignment— but it allows agents to make more informed choices. Motivated by the question of whether permitting agents to rely on local information can yield more efficient behavior, we focus on the second option.

Agent i estimates a set of possible joint AOHs $\vec{\Theta}_1^i$, and consequently a set of possible joint beliefs, $\mathcal{B}_1^i$, along with their probabilities of occurrence $P_1^i = \{\Pr(\vec{\theta}) \mid \vec{\theta} \in \vec{\Theta}_1^i\}$, by applying $\tau(\cdot)$ to all joint AOHs consistent with agent i's local AOH. Algorithm 1 describes, starting at time $t-1$ from the sets $\vec{\Theta}_{t-1}^i$, $\mathcal{B}_{t-1}^i$ and P_{t-1}^i, how to advance them to time t for a given joint action a_{t-1} and a local observation $o_{i,t}$. These sets serve as the basis for agent i to reason about action selection at the next step.

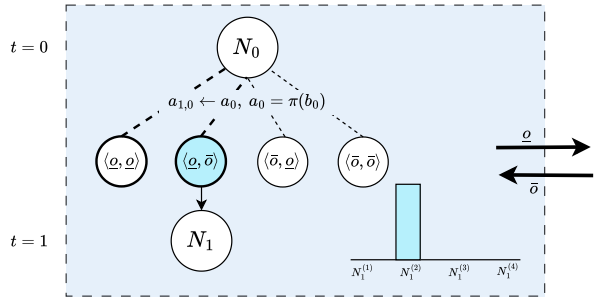
At each step, given a joint action, the belief space grows according to the number of joint observations compatible with agent i's local observation, which in turn depends on the size of the individual agents' observation spaces and

[3]For $h = \infty$ problems, the expected cumulative discounted reward.

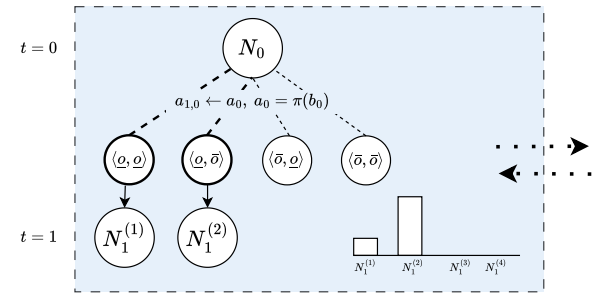
the number of agents. In the absence of communication, this space expands exponentially as the number of steps increases. The exponential growth poses a significant challenge for the planning problem, ultimately making direct enumeration of AOHs infeasible in practice. Moreover, when agents rely solely on local information, their belief sets gradually diverge, increasing the risk of poor system performance owing to inconsistencies in the agents' views. The effect of communication is precisely to collapse these sets and compute a single joint belief for the system, restoring a common view for all agents.

Algorithm 1 Belief Space Propagation for Agent i

function EXPAND($\vec{\Theta}^i_{t-1}, \mathcal{B}^i_{t-1}, P^i_{t-1}, a_{t-1}, o_{i,t}$)
$\vec{\Theta}^i_t = \emptyset$, $\mathcal{B}^i_t = \emptyset$, $P^i_t = \emptyset$
for $\vec{\theta}_{t-1} \in \vec{\Theta}^i_{t-1}$ **do**
  Select b_{t-1} from $\mathcal{B}^i_{t-1}$ associated to $\vec{\theta}_{t-1}$
  Select $\Pr(\vec{\theta}_{t-1})$ from P^i_{t-1} associated to $\vec{\theta}_{t-1}$
  for $o_{-i} \in O_{-i} = \times_{j \in I,\ j \neq i} O_j$ **do**
    $\vec{\theta}_t = (\vec{\theta}_{t-1}, a_{t-1}, \langle o_{i,t}, o_{-i} \rangle)$
    $b_t = \tau(b_{t-1}, a_{t-1}, \langle o_{i,t}, o_{-i} \rangle)$
    $\Pr(\vec{\theta}_t) = \Pr(\vec{\theta}_{t-1}) \Pr(\langle o_{i,t}, o_{-i} \rangle | a_{t-1}, b_{t-1})$
    $\vec{\Theta}^i_t \leftarrow \vec{\theta}_t$, $\mathcal{B}^i_t \leftarrow b_t$, $P^i_t \leftarrow \Pr(\vec{\theta}_t)$
  end for
end for
return $\vec{\Theta}^i_t,\ \mathcal{B}^i_t,\ P^i_t$



(a) With communication



(b) Without communication

Fig. 3. Belief Space Propagation for the first time step, with and without communication, for a team of two agents. The perspective is that of Agent 1, executing local action $a_{1,0}$ and receiving local observation $o_{1,1} = \underline{o}$.

C. Solving the planning problem

Solving Dec-POMDPs with general communication yields a truly decentralized policy, which each agent can execute independently using only local information. However, this approach is computationally demanding, posing significant challenges for both exact and approximate solution methods [12]. By assuming free and instantaneous communication, the multi-agent planning problem is reduced to an MPOMDP, whose complexity is PSPACE [3], compared to the NEXP complexity[4] that arises when trying to solve a Dec-POMDP with general communication [13]. This reduction allows the use of state-of-the-art POMDP solvers, such as [14]–[16], to efficiently compute a *centralized policy* for the team. While the computational complexity of offline planning is significantly reduced, the main drawback is the need for continuous and potentially costly information exchange during execution, which may be impractical in real-world scenarios with communication constraints.

The present work's central tenet is to leverage inter-agent communication to execute the centralized policy in a decentralized manner. While the approach does not yield the optimal solution to the Dec-POMDP, it can provide high-quality solutions to a large family of problems where coordination between agents is only required at critical junctures—providing a method to identify such junctures, and communicate only then. We provide a strategy in which each agent integrates local information, reasons about other agents' histories to construct a joint belief set compatible with local information, and applies a decision rule for selecting locally optimal actions. For such a strategy, the crux is then determining whether acquiring additional information from other agents could increase the overall system reward, avoiding penalties that result from miscoordination.

III. Reasoning about Communication

Suppose that a centralized policy π for the underlying MPOMDP has been generated during the offline planning phase. In this section, we present our proposed approach and examine three key aspects: (i) how agent i selects a local optimal joint action from the set of estimated joint beliefs given the centralized policy, (ii) how this selection relates to the optimal joint actions identified by the other agents, and (iii) under which conditions communication is unnecessary at a given time step.

A. Identification of the optimal action: local perspective

Agent i must reason locally to select the optimal joint action a^i_t and subsequently derive from it the local action $a^i_{i,t}$ to execute at time step t. A straightforward approach is to select a single representative belief $\hat{b}$, such as the mean belief or the most likely belief, and compute the action from the centralized policy, $a^i_t = \pi(\hat{b})$. Alternatively, the agent may adopt a risk-averse strategy by selecting the action that maximizes the worst-case value over the belief set, providing safety guarantees in high-risk settings. However, while the

[4]Complexity analysis is valid for finite-horizon problems.

first approach carries the risk of suboptimal decisions when the belief pool is highly diverse, the second tends to produce overly conservative behavior.

We propose an approach that overcomes these limitations by evaluating the expected reward across the distribution of possible joint beliefs and selecting the action that maximizes this expectation. In particular, the Q-POMDP heuristic is introduced to determine the joint action maximizing the expected reward over the distribution of possible joint beliefs if the agents were to communicate their AOHs in the next time step and all future time steps. It is based on the calculation of the *one-step lookahead value function* Q that defines the expected reward for taking a joint action a in the current time step, at belief b_t, and acting according to the centralized policy in the future:

$$Q(b_t, a) = \sum_{s \in S} R(s, a) b_t(s) + \gamma \sum_{o \in O} \Pr(o|a, b_t) V^{\pi}(\tau(b_t, a, o)).$$

For each agent i, the local optimal joint action without communication, $a^i_{t,NC}$, is selected by maximizing the expected reward over the distribution of all possible posterior joint beliefs:

$$a^i_{t,NC} = \arg\max_{a \in A} \sum_{\vec{\theta}_t \in \vec{\Theta}^i_t} \Pr(\vec{\theta}_t) Q(b_t(\vec{\theta}_t), a), \tag{1}$$

with $b_t(\vec{\theta}_t) \in \mathcal{B}^i_t$ joint belief associated with the joint AOH $\vec{\theta}_t$. The expected value for the system evaluated by agent i, given by the execution of a^i_{NC}, is:

$$V^i_{t,NC} = \sum_{\vec{\theta}_t \in \vec{\Theta}^i_t} \Pr(\vec{\theta}_t) Q(b_t(\vec{\theta}_t), a^i_{t,NC}). \tag{2}$$

If the estimated joint belief sets are identical across all agents, all agents deterministically select the same optimal joint action. However, given each agent's local perspective, this will not always be the case.

B. Identification of the optimal action: global perspective

Let us analyze the relationship between the joint action selected by agent i and those selected by the other agents, given that each acts locally based on its own belief pool while sharing the same centralized policy, decision rule, and common knowledge. A possible approach is the robust strategy, in which agent i assumes no knowledge about the other agents' AOH and expands its estimated joint belief set considering all possible actions and observations of the other agents. In this case, the QPOMDP heuristic selects an action that is "good" across all possible behaviors of the other agent, avoiding overconfidence but potentially limiting performance due to its conservatism. Conversely, agent i can assume that the other agents will behave according to its identified local optimal joint action, effectively narrowing the set of possible AOHs. This can provide computational advantages and higher rewards when the belief pools align, but it also introduces a significant risk of failure if the assumptions about the other agents' behavior are incorrect. Between the fully uncertain view and the optimistic view, intermediate approaches can leverage an expected-utility decision rule over the local belief pool while considering a restricted subset of plausible actions for the other agents, rather than marginalizing over the entire action space or assuming a single predicted action. These approaches are more optimistic than fully robust marginalization, yet more resilient than assuming perfect knowledge of the other agents' behavior. However, they may require evaluating multiple candidate actions, thereby increasing the computational complexity and extending the processing time.

In this paper, we examine the implications of adopting the optimistic approach, motivated by the observation that in domains characterized by high agent independence, precise step-by-step knowledge of other agents' behaviors is not essential for mission execution.

C. When to communicate: proposed approach

At each time step, agent i evaluates the *Value of Information* (VOI), which quantifies the expected benefit of receiving information from other agents relative to the cost of communication. There is no unique way to calculate this benefit. Information-theoretic methods measure the expected information gain from other agents' observations, normalized by communication cost. Regret-based heuristics consider the potential loss incurred by acting without communication, while risk- or safety-oriented methods focus on avoiding low-reward outcomes in the worst case. Sensitivity-based heuristics, in turn, evaluate how strongly an agent's expected reward depends on the actions of others.

In this paper, we define the VOI as the expected gain obtained by accessing the observations of other agents at the current time step and in the future, relative to the cost of communication. By measuring VOI in terms of expected return, the communication decision is consistent with the same optimization criterion used for action selection without communication. Moreover, this formulation enables computation using the existing belief representation, avoiding the need for additional computations or manually tuned thresholds. Leveraging the QPOMDP heuristic, we estimate the VOI as:

$$V^i_{t,C} = \sum_{\vec{\theta}_t \in \vec{\Theta}^i_t} \Pr(\vec{\theta}_t) \max_{a \in A} \Big[Q(b_t(\vec{\theta}_t), a) \Big]. \tag{3}$$

Communication is considered advantageous if the expected value of the received information from other agents exceeds the cost of communication:

$$|V^i_{t,C} - V^i_{t,NC}| \geq C_{\text{comm}}.$$

Algorithm 2 describes the execution phase from the perspective of agent i, highlighting the processes of belief space propagation, VOI evaluation, and optimal action selection. Before execution, agent i initializes its set of possible joint beliefs with the initial state distribution, which represents common knowledge across all agents (line 1). It further initializes variables to record its local AOH (line 2), and

Algorithm 2 Communication planning for Agent i

1: Initialize $\vec{\Theta}_0^i = \emptyset$, $\mathcal{B}_0^i = b_0$, $P_0^i = 1.0$
2: Initialize local AOH $\vec{\theta}_{i,0} = \emptyset$
3: Initialize $\mathcal{C} = \langle b_{t_c} = b_0, \vec{\theta}_{t_c} = \emptyset, t_c = 0 \rangle$
4: $\tau_{\text{ask}} = \text{false}$, $\tau_{\text{tell}} = \text{false}$
5: $a_0 \leftarrow \pi(b_0)$
6: **for** $t = 1, 2, \ldots, h$ **do**
7: Execute local action $a_{i,t-1}$
8: Receive local observation $o_{i,t}$
9: $\vec{\theta}_{i,t} = (\vec{\theta}_{i,t-1}, a_{i,t-1}, o_{i,t})$
10: $\vec{\Theta}_t^i, \mathcal{B}_t^i, P_t^i \leftarrow \text{EXPAND}(\vec{\Theta}_{t-1}^i, \mathcal{B}_{t-1}^i, P_{t-1}^i, a_{t-1}, o_{i,t})$
11: Calculate $a_{t,NC}^i$ from (1) and $V_{t,NC}^i$ from (2)
12: Calculate $V_{t,C}^i$ from (3)
13: **if** $|V_{t,C}^i - V_{t,NC}^i| \geq C_{\text{comm}}$ **then**
14: $\tau_{\text{ask}} = \text{true}$
15: Ask for other agents' AOH $\vec{\theta}_{-i,t}$
16: $b_t, \vec{\theta}_t \leftarrow \text{RESTORE}(\mathcal{C}, \vec{\theta}_{i,t}, \vec{\theta}_{-i,t}, t)$
17: $b_{t_c} \leftarrow b_t$, $\vec{\theta}_{t_c} = \vec{\theta}_t$, $t_c \leftarrow t$
18: Calculate $a_t \leftarrow \pi(b_t)$
19: $\vec{\Theta}_t^i = \vec{\theta}_t$, $\mathcal{B}_t^i = b_t$, $P_t^i = 1.0$
20: **else**
21: Check request for Communication
22: **if** $\tau_{\text{tell}} = \text{true}$ **then**
23: Send $\vec{\theta}_{i,t}$
24: **end if**
25: $a_t \leftarrow a_{t,NC}^i$
26: **end if**
27: $\tau_{\text{ask}} = \text{false}$, $\tau_{\text{tell}} = \text{false}$
28: **end for**

Algorithm 3 Joint Belief calculation from last time step of communication for Agent i

1: **function** RESTORE($\mathcal{C}, \vec{\theta}_{i,t}, \vec{\theta}_{-i,t}, t$)
2: $b_{t_c}, \vec{\theta}_{t_c}, t_c \leftarrow \mathcal{C}$
3: **for** $k = t_c$ **to** $t - 1$ **do**
4: Extract a_i and o_i from $\vec{\theta}_{i,k}$
5: Extract a_{-i} and o_{-i} from $\vec{\theta}_{-i,k}$
6: $\vec{\theta}_{k+1} = (\vec{\theta}_k, \langle a_i, a_{-i} \rangle, \langle o_i, o_{-i} \rangle)$
7: $b_{k+1} = \tau(b_k, \langle a_i, a_{-i} \rangle, \langle o_i, o_{-i} \rangle)$
8: **end for**
9: **return** $b_t, \vec{\theta}_t$

the information received at the last communication event, i.e., (i) the most recent joint belief reconstructed through communication, (ii) the corresponding joint AOH, and (iii) the time step at which communication occurred (line 3). At line 4, agent i initializes two binary flags: the first, τ_{ask}, is true if agent i requires information from other agents, and the second, τ_{tell}, is true if other agents request information from agent i.

The execution phase starts with each agent executing its local action, obtained from the joint action specified by the centralized policy π. This action is *coordinated*, in the sense that it is consistent across all agents and follows from their common knowledge of the policy and the initial system state distribution. Upon receiving a local observation, agent i updates its local AOH and constructs the set of possible joint histories, joint beliefs, and their associated probabilities with the propagation function (lines 9–10). It then computes the action and the value that would be obtained by acting solely on local information, comparing it with the expected value attainable if the observations of other agents were known. Communication is considered beneficial if the expected improvement exceeds the communication cost.

If this condition holds, agent i requests the local AOHs of the other agents, reconstructs the current joint belief using Algorithm 3, and selects the joint action prescribed by the centralized policy. Otherwise, it checks whether communication has been requested by other agents (line 22). If so, it transmits its local AOH; if not, it executes the action computed under the assumption of no communication.

In many practical domains, exact VOI computation can be prohibitively expensive, even under a myopic formulation. When the space of possible joint beliefs is large, sampling-based Monte Carlo methods can be employed to improve computational tractability and flexibility, at the cost of approximation error and potential suboptimality.

IV. Experiments

We evaluate the proposed algorithm and compare it with state-of-the-art methods through a series of experiments on several Dec-POMDP benchmarks. Specifically, we assess the planning solutions produced by the ALWAYS-COMM strategy (**M-1**), corresponding to the MPOMDP solution, an algorithm with the same logic as the ACE-PJB-Comm algorithm [17] (**M-2**), and our proposed approach (**M-3**). For all cases, we first compute an offline centralized policy using the SARSOP algorithm [14], and then execute it at runtime, measuring performance across a Monte Carlo (MC) simulation of N runs, across the following metrics:

- average cumulative reward r over the MC runs (with 95% confidence intervals);
- average frequency of communication f_c (%);
- average computational time t per run, averaged over MC runs, required to decide whether to communicate and what action to take next;
- maximum dimension of the set of estimated joint AOHs over which each algorithm has reasoned to decide the optimal joint action (i.e., the maximum dimension of the joint belief set $\mathcal{B}_t^i$ before a decision on communication has been made), averaged across the MC runs $\dim(\vec{\Theta}')$;
- average frequency of disagreement about the joint action $f_{a_{\neq}}$ (%) (i.e., the frequency with which the optimal joint action locally selected by an agent differs from that of the others), a measure of the *action inconsistency*.

Computational times were measured using a demonstrative, non-optimized implementation, with utilities from the *POMDPs.jl* library [18]. All experiments were run on a PC with an Intel(R) Core(TM) Ultra 7 CPU with 32.0 GB RAM. They are reported solely to illustrate relative differences in runtime across strategies and are not intended as performance benchmarks. For compactness, we denote each method–cost pair as **M-**$x_{C_{\text{comm}}}$, where **M-**x indicates the method ($x = 1, 2, 3$) and C_{comm} the communication cost. Benchmark problems include *Multi-Agent Tiger Problem* [19], *Broadcast Channel* [20], *Meeting in a Grid 2×2* [21], and *Mars Rovers* [22], with their parameters summarized in Table

TABLE I

BENCHMARK PROBLEM SPECIFICATIONS

Problem	n	$\|S\|$	$\|A\|$	$\|O\|$	max_steps	N
Multi-Agent Tiger	2	2	9	4	6	100
Broadcast Channel	2	4	4	25	6	100
Meeting in a Grid	2	16	25	16	5	15
Mars Rovers	2	256	36	64	15	3

I.[5] To prevent the estimated joint belief sets from growing excessively, a pruning threshold of 10^{-4} is introduced.

A. Results

Simulation results are presented in Table II for the *Multi-Agent Tiger Problem*, Table III for the *Broadcast Channel*, Table IV for *Meeting in a Grid 2×2*, and Table V for *Mars Rovers*. In domains such as the *Multi-Agent Tiger Problem*, coordination between agents is crucial, as significant penalties are incurred if they fail to select the same action at each decision point. Our algorithm often opts to communicate because the risk associated with miscoordination is substantial. Even when local information is available, the danger of acting based purely on individual observations remains significant. Therefore, the agents prefer to incur the cost of communication rather than act independently and risk a misaligned decision, up to the point where communication becomes prohibitively expensive. At this stage, whereas a conservative approach, such as **M-2**, continues to listen indefinitely, in our algorithm, agents may choose to open a door. This can result in either high or very low rewards depending on the simulation outcome, causing a substantially higher variance in the observed returns. The same applies to the results for the *Broadcast Channel Problem*.

In domains where agents operate with partial autonomy in decision-making, given by weakly coupled rewards, or with less severe consequences for acting in disagreement, our algorithm proves effective. By integrating local information, agents estimate a joint belief set that more accurately reflects the current joint state, improving the evaluation of VOI and the quality of the individual decision-making process. In the *Meeting in a Grid* and in the *Mars Rovers* problems, agents can identify highly rewarded actions without communicating at every time step. As shown in the last column of Table IV and Table V, even without enforced coordination as in **M-2** algorithm, agents either (i) select the same joint action and act in coordination solely by reasoning about the behavior and observations of the other agents, or (ii) judge that acting without coordination is not risky, and the high rewards confirm the effectiveness of their choice.

Beyond rewards, we also analyze computational aspects, specifically computational time and the maximum size of the joint belief sets that each agent must evaluate to select the optimal action. These factors are closely related: as the number of elements in the joint belief set grows, computational demands increase rapidly, eventually rendering real-time execution infeasible. Prior work has attempted to address this challenge by constraining communication frequency—e.g., requiring agents to exchange information at least once every K steps [5]— or by pruning and clustering histories to control memory growth [10], [23]. In contrast, by integrating locally available information and leveraging the optimistic approach to joint action selection, our algorithm significantly reduces the computational effort required for belief set management.

[5]The notation $|X|$ is used to denote the dimension of the space X.

TABLE II

PERFORMANCE FOR THE MULTI-AGENT TIGER PROBLEM WITH $n = 2$.

Method	r	f_c (%)	t (s)	$\dim(\vec{\Theta}')$	$f_{a_{\neq}}$ (%)
M-1$_{0.0}$	$\mathbf{7.460 \pm 4.673}$	100.0	0.000	1.0	0.0
M-2$_{0.0}$	6.560 ± 4.223	44.5	0.029	131.8	0.0
M-3$_{0.0}$	$\mathbf{7.460 \pm 4.673}$	100.0	0.013	1.0	0.0
M-1$_{0.2}$	$\mathbf{7.116 \pm 4.678}$	100.0	0.000	1.0	0.0
M-2$_{0.2}$	6.348 ± 4.223	44.5	0.008	131.8	0.0
M-3$_{0.2}$	$\mathbf{7.116 \pm 4.678}$	100.0	0.000	1.0	0.0
M-1$_{0.8}$	$\mathbf{6.084 \pm 4.696}$	100.0	0.000	1.0	0.0
M-2$_{0.8}$	5.084 ± 3.287	25.27	0.031	491.08	0.0
M-3$_{0.8}$	$\mathbf{6.084 \pm 4.696}$	100.0	0.000	1.0	0.0
M-1$_{2.0}$	4.020 ± 4.740	100.0	0.000	1.0	0.0
M-2$_{2.0}$	$\mathbf{4.040 \pm 3.256}$	25.27	0.031	491.08	0.0
M-3$_{2.0}$	-3.850 ± 6.546	47.5	0.000	3.3	15.97
M-1$_{10.0}$	$\mathbf{-9.740 \pm 5.297}$	100.0	0.000	1.0	0.0
M-2$_{10.0}$	-12.000 ± 0.000	0.00	0.135	3172.0	0.0
M-3$_{10.0}$	-12.890 ± 8.661	0.00	0.000	9.16	19.50

TABLE III

PERFORMANCE FOR THE BROADCAST CHANNEL PROBLEM.

Method	r	f_c (%)	t (s)	$\dim(\vec{\Theta}')$	$f_{a_{\neq}}$ (%)
M-1$_{0.0}$	$\mathbf{5.690 \pm 0.11}$	100.0	0.000	1.0	0.0
M-2$_{0.0}$	$\mathbf{5.690 \pm 0.11}$	8.8	0.104	137.3	0.0
M-3$_{0.0}$	$\mathbf{5.690 \pm 0.11}$	24.8	0.008	27.68	5.17
M-1$_{0.05}$	5.440 ± 0.110	100.0	0.000	1.0	0.0
M-2$_{0.05}$	$\mathbf{5.668 \pm 0.115}$	8.8	0.085	137.3	0.0
M-3$_{0.05}$	5.628 ± 0.107	24.8	0.001	27.68	5.17
M-1$_{0.2}$	4.690 ± 0.11	100.0	0.000	1.0	0.0
M-2$_{0.2}$	$\mathbf{5.608 \pm 0.118}$	3.2	0.106	151.9	0.0
M-3$_{0.2}$	5.560 ± 0.119	0.00	0.002	40.15	7.33
M-1$_{0.4}$	3.690 ± 0.11	100.0	0.000	1.0	0.0
M-2$_{0.4}$	$\mathbf{5.560 \pm 0.119}$	0.00	0.118	174.0	0.0
M-3$_{0.4}$	$\mathbf{5.560 \pm 0.119}$	0.00	0.002	40.15	7.33

V. CONCLUSION

This paper contributes a method for multi-agent systems operating under uncertainty, where broadcast communication is feasible but costly. Agents are able to make effective use of a centralized policy, despite their local information generally being insufficient to reconstruct the system's global joint belief. Instead, the agents each integrate local observations and predict other agents' histories, generating a pool of estimated joint beliefs. Each individual agent's approach to action selection is then to maximize the estimated average reward. So doing provides a basis for predicting outcomes realized in the absence of agent-interaction, and gives a basis for comparison against performance expected after information exchange. Weighing the value expected to be

TABLE IV

PERFORMANCE FOR THE MEETING IN A GRID 2×2 PROBLEM.

Method	r	f_c (%)	t (s)	$\dim(\vec{\Theta}')$	$f_{a_{\neq}}$ (%)
M-1$_{0.0}$	$\mathbf{3.333 \pm 0.563}$	100.0	0.004	1.0	0.0
M-2$_{0.0}$	$\mathbf{3.333 \pm 0.563}$	50.00	0.235	2.2	0.0
M-3$_{0.0}$	$\mathbf{3.333 \pm 0.563}$	56.67	0.025	1.0	0.0
M-1$_{0.2}$	2.533 ± 0.563	100.0	0.000	1.0	0.0
M-2$_{0.2}$	2.480 ± 0.687	48.33	0.411	196.87	0.0
M-3$_{0.2}$	$\mathbf{2.680 \pm 0.783}$	56.67	0.002	1.8	6.67
M-1$_{0.4}$	1.733 ± 0.563	100.0	0.001	1.0	0.0
M-2$_{0.4}$	1.440 ± 0.626	18.33	2.137	905.4	0.0
M-3$_{0.4}$	$\mathbf{2.400 \pm 0.953}$	0.00	0.034	278.0	80.0
M-1$_{0.8}$	0.133 ± 0.563	100.0	0.001	1.0	0.0
M-2$_{0.8}$	1.733 ± 0.619	0.00	3.227	1396.0	0.0
M-3$_{0.8}$	$\mathbf{2.400 \pm 0.953}$	0.00	0.037	278.0	80.0

TABLE V

PERFORMANCE FOR THE MARS ROVERS PROBLEM.

Method	r	f_c (%)	t (s)	$\dim(\vec{\Theta}')$	$f_{a_{\neq}}$ (%)
M-1$_{0.0}$	$\mathbf{12.400 \pm 4.704}$	100.0	0.004	1.0	0.0
M-2$_{0.0}$	$\mathbf{12.400 \pm 4.704}$	44.44	8.166	2.0	0.0
M-3$_{0.0}$	$\mathbf{12.400 \pm 4.704}$	27.78	0.026	1.0	0.0
M-1$_{0.2}$	11.467 ± 4.181	100.0	0.003	1.0	0.0
M-2$_{0.2}$	12.133 ± 4.835	44.44	7.871	2.0	0.0
M-3$_{0.2}$	$\mathbf{12.200 \pm 4.704}$	27.78	0.017	1.0	0.0
M-1$_{0.4}$	10.533 ± 3.659	100.0	0.002	1.0	0.0
M-2$_{0.4}$	11.600 ± 4.704	55.56	8.584	4.0	0.0
M-3$_{0.4}$	$\mathbf{12.000 \pm 4.704}$	27.78	0.016	1.0	0.0
M-1$_{0.8}$	8.667 ± 2.613	100.0	0.003	1.0	0.0
M-2$_{0.8}$	10.800 ± 4.704	55.56	9.370	4.0	0.0
M-3$_{0.8}$	$\mathbf{11.600 \pm 4.704}$	27.78	0.022	2.0	0.0
M-1$_{2.5}$	0.733 ± 1.829	100.0	0.002	1.0	0.0
M-2$_{2.5}$	0.967 ± 6.207	6.67	12.129	36.0	0.0
M-3$_{2.5}$	$\mathbf{9.900 \pm 4.704}$	27.78	2.078	4.0	11.11
M-1$_{10.0}$	-34.267 ± 21.429	100.0	0.003	1.0	0.0
M-2$_{10.0}$	-1.533 ± 1.307	0.00	14.424	46.0	0.0
M-3$_{10.0}$	$\mathbf{1.067 \pm 4.835}$	0.00	1.653	4.0	33.33

gained by communication against its cost yields a rule for deciding when communication should occur.

The method described depends upon making predictions in the absence of real knowledge about what other agents are sensing. As such, its performance depends upon the risks entailed in having discrepant actions as a consequence of divergent beliefs. However, as our simulation results show, several benchmark coordination problems can tolerate some misalignment, and being conservative by only acting on jointly known data often sacrifices performance unduly.

REFERENCES

[1] Y. Rizk, M. Awad, and E. W. Tunstel, "Decision Making in Multiagent Systems: A Survey," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 10, no. 3, pp. 514–529, 2018.

[2] F. Oliehoek and C. Amato, *A Concise Introduction to Decentralized POMDPs*. Springer, 2016.

[3] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra, "Planning and acting in partially observable stochastic domains," *Artificial Intelligence*, vol. 101, no. 1, pp. 99–134, 1998.

[4] C. Goldman and S. Zilberstein, "Optimizing information exchange in cooperative multi-agent systems," *Proceedings of the International Conference on Autonomous Agents*, vol. 2, Jun. 2003.

[5] R. Nair, M. Roth, and M. Yokoo, "Communication for Improving Policy Computation in Distributed POMDPs," in *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems*, vol. 3, 2004, pp. 1098–1105.

[6] F. Wu, S. Zilberstein, and X. Chen, "Online planning for multi-agent systems with bounded communication," *Artificial Intelligence*, vol. 175, no. 2, pp. 487–511, 2011.

[7] P. J. Gmytrasiewicz and P. Doshi, "A Framework for Sequential Planning in Multi-Agent Settings," *Journal of Artificial Intelligence Research*, vol. 24, p. 49–79, Jul. 2005.

[8] P. Gmytrasiewicz, "How to Do Things with Words: A Bayesian Approach," *Journal of Artificial Intelligence Research*, vol. 68, pp. 753–776, Jul. 2020.

[9] L. Zhou and J. Luo, "A communication model for interactive POMDPs," in *Proceedings of the International Conference on Computer Science & Education*, 2012, pp. 169–174.

[10] R. Emery-Montemerlo, "Game-Theoretic Control for Robot Teams," PhD dissertation, Carnegie Mellon University, 2005.

[11] M. Roth, R. Simmons, and M. Veloso, "Reasoning about joint beliefs for execution-time communication decisions," in *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems*, 2005, pp. 786–793.

[12] S. Seuken and S. Zilberstein, "Formal models and algorithms for decentralized decision making under uncertainty," *Autonomous Agents and Multi-Agent Systems*, vol. 17, no. 2, pp. 190–250, Oct. 2008.

[13] C. Goldman and S. Zilberstein, "Decentralized Control of Cooperative Systems: Categorization and Complexity Analysis," *Journal of Artificial Intelligence Research*, vol. 22, Jun. 2011.

[14] H. Kurniawati, D. Hsu, and W. S. Lee, "SARSOP: Efficient Point-Based POMDP Planning by Approximating Optimally Reachable Belief Spaces," in *Proceedings of Robotics: Science and Systems*. MIT Press, 2008, pp. 65–72.

[15] G. Shani, J. Pineau, and R. Kaplow, "A survey of point-based POMDP solvers," *Autonomous Agents and Multi-Agent Systems*, vol. 27, pp. 1–51, 2012.

[16] M. Spaan and N. Vlassis, "Perseus: Randomized Point-based Value Iteration for POMDPs," *Journal of Artificial Intelligence Research*, vol. 24, 11 2004.

[17] M. Roth, "Execution-time communication decisions for coordination of multi-agent teams," Ph.D. dissertation, Carnegie Mellon University, USA, 2007, CMU-RI-TR-08-04.

[18] M. Egorov, Z. N. Sunberg, E. Balaban, T. A. Wheeler, J. K. Gupta, and M. J. Kochenderfer, "POMDPs.jl: A framework for sequential decision making under uncertainty," *Journal of Machine Learning Research*, vol. 18, no. 26, pp. 1–5, 2017.

[19] R. Nair, M. Tambe, M. Yokoo, D. Pynadath, and S. Marsella, "Taming decentralized POMDPs: towards efficient policy computation for multiagent settings," in *Proceedings of the International Joint Conference on Artificial Intelligence*, 2003, pp. 705–711.

[20] J. Ooi and G. Wornell, "Decentralized control of a multiple access broadcast channel: performance bounds," in *Proceedings of the Conference on Decision and Control*, vol. 1, 1996, pp. 293–298.

[21] D. S. Bernstein, E. A. Hansen, and S. Zilberstein, "Bounded policy iteration for decentralized POMDPs," in *Proceedings of the International Joint Conference on Artificial Intelligence*, 2005, pp. 1287–1292.

[22] C. Amato and S. Zilberstein, "Achieving goals in decentralized POMDPs," in *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems*, 2009, pp. 593–600.

[23] J.-Y. Kwak, R. Yang, Z. Yin, M. E. Taylor, and M. Tambe, "Teamwork and coordination under model uncertainty in Dec-POMDPs," in *Proceedings of the AAAI Conference on Interactive Decision Theory and Game Theory*, 2010, pp. 30–36.