Optimal Communication and Control Strategies in a Cooperative Multiagent MDP Problem

Sagar Sudhakara, Dhruva Kartik, Rahul Jain, *Senior Member, IEEE*, and Ashutosh Nayyar, *Senior Member, IEEE*

***Abstract*—The problem of controlling cooperative multiagent systems under different models of information sharing among agents has received significant attention in the recent literature. In this article, we consider a setup where rather than committing to a fixed and nonadaptive information sharing protocol (e.g., periodic sharing or no sharing, etc.), agents can dynamically decide at each time step whether to share information with each other and incur the resulting communication cost. This setup requires a joint design of agents' communication and control strategies in order to optimize the tradeoff between communication costs and the control objective. We first show that agents can ignore a big part of their private information without compromising the system performance. We then provide a common-information-approach-based solution for the strategy optimization problem. This approach relies on constructing a fictitious partially observable markov decision process (POMDP) whose solution (obtained via a dynamic program) characterizes the optimal strategies for the agents. We extend our solution to incorporate time-varying packet-drop channels and constraints on when and how frequently agents can communicate.**

***Index Terms*—Agents and autonomous systems, cooperative control, multiagent systems, POMDP, stochastic optimal control.**

I. Introduction

The problem of sequential decision-making by a team of collaborative agents has received significant attention in the recent literature. The goal in such problems is to jointly design decision/control strategies for the multiple agents in order to optimize a performance metric for the team. The nature of this joint strategy optimization problem as well as the best achievable performance depend crucially on the *information structure* of the problem. Intuitively, the information structure of a multiagent problem specifies what information is available to each agent at each time. Depending on the underlying communication environment, a wide range of information structures can arise. If communication is costless and unrestricted, all agents can share all information with each other. If communication is too costly or physically impossible, agents may not be able to share any information at all. It could also be the case that agents can communicate only periodically or that the ability to communicate varies among the agents leading to one-directional communication between certain pairs of agents. Each of these communication models corresponds to a different information structure, which, in turn, specifies the class of feasible decision/control strategies for the agents.

In this article, we consider a setup where rather than committing to a fixed and nonadaptive information sharing protocol (e.g., periodic sharing or no sharing, etc.), agents can dynamically decide at each time step whether to share information with each other and incur the resulting communication cost. Thus, at each time step, agents have to make two kinds of decisions—1) communication decisions that govern the information sharing and 2) control decisions that govern the evolution of the agents' states. The two kinds of agents' strategies—1) communication strategies and 2) control strategies—need to be jointly designed in order to optimize the tradeoff between communication costs and the control objective.

Related Work: There is a significant body of prior work on decentralized control and decision-making in multiagent systems. We focus on works where the dynamic system can be viewed as a Markov chain jointly being controlled by multiple agents/controllers. We can organize this literature based on the underlying information structure (or the information sharing protocol).

In decentralized Markov decision processes (Dec-MDPs) and decentralized partially observable Markov decision processes (Dec-partially observable markov decision process (POMDP)), each agent receives a partial or noisy observation of the current system state [1]. These agents cannot communicate or share their observations with each other and can only use their private action-observation history to select their control actions. Several methods for solving such generic Dec-POMDPs exist in the literature [2], [3], [4], [5]. However, these generic methods either involve prohibitively large amount of computation or cannot guarantee optimality. For certain Dec-MDPs and Dec-POMDPs with an additional structure, such as transition independence in Dec-MDPs [6], [7] or one-sided information sharing [8], one can derive additional structural properties of the optimal strategy and use these properties to make the computation more tractable.

In the decentralized stochastic control literature, a variety of information structures (obtained from different information sharing protocols) have been considered [9], [10], [11]. For example, Nayyar et al. [9] consider the case where agents share their information with each other with a fixed delay. The work in [10] provides a unified treatment for a range of information sharing protocols including periodic sharing, sharing of only control actions, etc. The authors in [11] and [12] consider a setup where only the agents' actions are shared with others.

In emergent communication, agents have access to a cheap talk channel, which can be used for communication. The authors in [13], [14], and [15] propose methods for jointly learning the control and communication strategies in such settings. The key communication issue in these works is to design the most effective way of encoding the available information into the communication alphabet [16]. In contrast, the communication issue in our setup is whether the cost of sharing states is worth the potential control benefit.

In multiagent actor–critic literature, multiagent deep deterministic policy gradient (MADDPG) method [17] uses a dedicated centralized critic for each agent in semicompetitive domains, demonstrating compelling empirical results in continuous action environments. Foerster et al. [18] propose a new multiagent actor–critic method called counterfactual multiagent (COMA) policy gradients. COMA uses a centralized critic to estimate the Q-function and decentralized actors to optimize the agents' policies. MADDPG and COMA methods in [17] and [18] provide practically implementable heuristics but do not have optimality guarantees.

Manuscript received 17 January 2023; revised 19 January 2023 and 10 November 2023; accepted 2 April 2024. Date of publication 8 April 2024; date of current version 27 September 2024. The work was supported by NSF under Grant ECCS 2025732 and Grant ECCS 1750041. Recommended by Associate Editor N. Li. *(Corresponding author: Sagar Sudhakara.)*

The authors are with the Department of Electrical Engineering, University of Southern California, Los Angeles, CA 90089 USA (e-mail: sagarsud@usc.edu; mokhasun@usc.edu; ashutosn@usc.edu).

Digital Object Identifier 10.1109/TAC.2024.3386454

1558-2523 © 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: University of Southern California. Downloaded on October 16,2024 at 18:46:09 UTC from IEEE Xplore. Restrictions apply.

In our model, agents at each time make an explicit choice regarding sharing their information with each other. We seek to jointly design this information sharing strategy and the agents' control strategies. This problem and many of the problems considered in the prior literature can be reduced to Dec-POMDPs by a suitable redefinition of states, observations, and actions. However, as demonstrated in [8], a generic Dec-POMDP-based approach for problems with (limited) interagent communication involves a very large amount of computation since it ignores the underlying communication structure. Instead, we derive some structural properties of the strategies that significantly simplify the strategy design. We then provide a dynamic-program-based solution using the common information approach. To the best of our knowledge, our information sharing mechanism has not been analyzed before.

Contributions:

1) We first show that agents can ignore a big part of their private information without compromising the system performance. This is done by using an agent-by-agent argument where we fix the strategies of one agent arbitrarily and find a sufficient statistic for the other agent. This sufficient statistic turns out to be a subset of the agent's private information. This reduction in private information narrows down the search for optimal strategies to a class of simpler strategies.
2) We then adopt the common-information-based solution approach for finding the optimal strategies. This approach relies on constructing an equivalent POMDP from the perspective of a fictitious coordinator that knows the common information among the agents. The solution of this POMDP (obtained via a dynamic program) characterizes the optimal strategies for the agents.
3) Finally, we extend our setup to incorporate time-varying packet-drop channels and constraints on when and how frequently agents can communicate with each other. We show that our solution approach can be easily modified to incorporate these features using a natural augmentation of the state in the coordinator's POMDP.

Notation: Random variables are denoted with upper case letters (X, Y, etc.), their realization with lower case letters (x, y, etc.), and their space of realizations by script letters ($\mathcal{X}$, $\mathcal{Y}$, etc.). Subscripts denote time and superscripts denote the subsystem; e.g., X^i_t denotes the state of subsystem i at time t. The short hand notation $X^i_{1:t}$ denotes the collection $(X^i_1, X^i_2, \dots, X^i_t)$. X_t denotes (X^1_t, X^2_t) and M_t denotes (M^1_t, M^2_t). $\triangle(\mathcal{X})$ denotes the probability simplex for the space $\mathcal{X}$. $\mathrm{P}(A)$ denotes the probability of an event A. $\mathbb{E}[X]$ denotes the expectation of a random variable X. $\mathbb{1}_A$ denotes the indicator function of event A. For simplicity of notation, we use $\mathrm{P}(x_{1:t}, u_{1:t-1})$ to denote $\mathrm{P}(X_{1:t} = x_{1:t}, U_{1:t-1} = u_{1:t-1})$ and a similar notation for conditional probability. We use $-i$ to denote agent/agents other than agent i.

II. Problem Formulation

Consider a discrete-time system with two agents. Let $X^i_t \in \mathcal{X}^i$ denote the local state of agent i for $i = 1, 2$. $X_t := (X^1_t, X^2_t)$ represents the local state of both agents. The initial local states, (X^1_1, X^2_1), of both agents are independent random variables with the initial local state X^i_1 having the probability distribution $P_{X^i_1}$. Each agent perfectly observes its own local state. Let $U^i_t \in \mathcal{U}^i$ denote the control action of agent i at time t and $U_t := (U^1_t, U^2_t)$ denote the control actions of both agents at time t. The local state of agent $i, i = 1, 2$, evolves according to

$$X^i_{t+1} = k^i_t(X^i_t, U^i_t, W^i_t) \tag{1}$$

where $W^i_t \in \mathcal{W}^i$ is the disturbance in dynamics with probability distribution P_{W^i}. The initial state X_1 and the disturbances $\{W^i_t\}_{t\geq 1}$, $i = 1, 2$, are independent random variables. Note that the next local state of agent i depends on the current local state and control action of agent i alone. The dynamics of the two agents are independent of each other.

In addition to deciding the control actions at each time, the two agents need to decide whether or not to initiate communication at each time. We use the binary variable $M^i_t \in \{0, 1\}$ to denote the communication decision taken by agent i. Let $M^{or}_t := \max(M^1_t, M^2_t)$ and let Z^{er}_t represent the information exchanged between the agents at time t. In our model, communication is initiated when any agent decides to communicate (i.e., $M^i_t = 1$) but agents may lose packets or fail to communicate with probability p_e. Based on the communication model described earlier, Z^{er}_t is given as

$$Z^{er}_t = \begin{cases} X^{1,2}_t, \text{ with probability } 1 - p_e & \text{if } M^{or}_t = 1. \\ \phi, \quad \text{with probability } \ p_e & \text{if } M^{or}_t = 1. \\ \phi, & \text{if } M^{or}_t = 0. \end{cases} \tag{2}$$

Information structure and decision strategies: At the beginning of the tth time step, the information available to agent i is given by

$$I^i_t = \{X^i_{1:t}, U^i_{1:t-1}, Z^{er}_{1:t-1}, M^{1,2}_{1:t-1}\}. \tag{3}$$

Agent i can use this information to make its communication decision at time t. Thus, M^i_t is chosen as a function of I^i_t according to $M^i_t = f^i_t(I^i_t)$, where the function f^i_t is referred to as the communication strategy of agent i at time t. After the communication decisions are made and the resulting communication (if any) takes place, the information available to agent i is

$$I^i_{t^+} = \{I^i_t, Z^{er}_t, M^{1,2}_t\}. \tag{4}$$

Agent i then chooses its control action according to $U^i_t = g^i_t(I^i_{t^+})$, where the function g^i_t is referred to as the control strategy of agent i at time t.

$f^i := (f^i_1, f^i_2, \dots, f^i_T)$ and $g^i := (g^i_1, g^i_2, \dots, g^i_T)$ are called the communication and control strategy of agent i, respectively.

Strategy optimization problem: At time t, the system incurs a cost $c_t(X^1_t, X^2_t, U^1_t, U^2_t)$ that depends on the local states and control actions of both agents. Whenever agents decide to share their states with each other, they incur a state-dependent communication cost $\rho(X_t)$. The communication cost $\rho(X_t)$ includes the energy cost involved in transmission and the computation cost involved in encoding and decoding messages. The system runs for a time horizon T. The objective is to find communication and control strategies for the two agents in order to minimize the expected value of the sum of control and communication costs over the time horizon T.

$$\mathbb{E}\left[\sum_{t=1}^{T} c_t(X_t, U_t) + \rho(X_t)\mathbb{1}_{\{M^{or}_t = 1\}}\right]. \tag{5}$$

Remark 1: Even though we are formulating the problem for two agents, it can be easily extended to n agents with the communication protocol that if any one agent initiates communication all agents broadcast their state. All key results in the article apply for n agents setup with minor adjustments in the proof.

III. Preliminary Results and Simplified Strategies

In this section, we show that agents can ignore parts of their information without losing optimality. This removal of information narrows the search for optimal strategies to a class of simpler strategies and is a key step in our approach for finding optimal strategies. To proceed, we first split the information available to the agents into two parts—1) common information (which is available to both agents) and 2) private information (which is everything except the common information).

1) At the beginning of time step t, before the communication decisions are made, the common information is defined as $C_t := (Z^{er}_{1:t-1}, M^{1,2}_{1:t-1})$.
2) After the communication decisions are made and the resulting communication (if any) takes place, the common information is defined as: $C_{t^+} = (Z^{er}_{1:t}, M^{1,2}_{1:t})$.

The following lemma establishes a key conditional independence property that will be critical for our analysis.

Lemma 1 (Conditional independence property): Consider any arbitrary choice of communication and control strategies for the two agents. Then, at any time t, the two agents' local states and control

Authorized licensed use limited to: University of Southern California. Downloaded on October 16,2024 at 18:46:09 UTC from IEEE Xplore. Restrictions apply.

actions are conditionally independent given the common information C_t (before communication) or C_{t+} (after communication). That is, if c_t, c_{t+} are the realizations of the common information before and after communication, respectively, then for any realization $x_{1:t}, u_{1:t-1}$ of states and actions, we have

$$\mathrm{P}(x_{1:t}, u_{1:t-1}|c_t) = \prod_{i=1}^{2} \mathrm{P}(x^i_{1:t}, u^i_{1:t-1}|c_t) \tag{6}$$

$$\mathrm{P}(x_{1:t}, u_{1:t}|c_{t+}) = \prod_{i=1}^{2} \mathrm{P}(x^i_{1:t}, u^i_{1:t}|c_{t+}). \tag{7}$$

Furthermore, $\mathrm{P}(x^i_{1:t}, u^i_{1:t-1}|c_t)$ and $\mathrm{P}(x^i_{1:t}, u^i_{1:t}|c_{t+})$ depend only on agent i' strategy and not on the strategy of agent $-i$.

Proof: See Appendix A. ■

The following proposition shows that agent i at time t can ignore its past states and actions, i.e., $X^i_{1:t-1}$ and $U^i_{1:t-1}$, without losing optimality. This allows agents to use simpler strategies where the communication and control decisions are functions only of the current state and the common information.

Proposition 1: Agent i, $i = 1, 2$, can restrict itself to strategies of the following form:

$$M^i_t = \bar{f}^i_t(X^i_t, C_t) \tag{8}$$

$$U^i_t = \bar{g}^i_t(X^i_t, C_{t+}) \tag{9}$$

without loss of optimality. In other words, at time t, agent i does not need the past local states and actions, $X^i_{1:t-1}, U^i_{1:t-1}$, for making optimal decisions.

Proof: To prove this result, we fix agent $-i$'s strategy to an arbitrary choice and then show that agent i's decision problem can be modeled as an MDP in a suitable state space. The result then follows from the fact that Markovian strategies are optimal in an MDP. See Appendix B for details. ■

IV. Centralized Reformulation Using Common Information

In this section, we provide a centralized reformulation of the multiagent strategy optimization problem using the common information approach in [10]. The main idea of the approach is to formulate an equivalent single-agent POMDP problem, solve the equivalent POMDP using a dynamic program, and then translate the results back to the original problem.

Because of Proposition 1, we will only consider strategies of the form given in (8) and (9). Following the approach in [10], we construct an equivalent problem by adopting the point of view of a fictitious coordinator that observes only the common information among the agents (i.e., the coordinator observes C_t before communication and C_{t+} after Z^{er}_t is realized) but not the current local states (i.e., $X^i_t, i = 1, 2$). Before communication at time t, the coordinator chooses a pair of *prescriptions*, $\Gamma_t := (\Gamma^1_t, \Gamma^2_t)$, where Γ^i_t is a mapping from X^i_t to M^i_t (more precisely, Γ^i_t maps $\mathcal{X}^i$ to $\{0, 1\}$). The interpretation of the prescription is that it is a directive to the agents about how they should use their local state information to make the communication decisions. Thus, agent i generates its communication decision by evaluating the function Γ^i_t on its current local state: $M^i_t = \Gamma^i_t(X^i_t)$. Similarly, after the communication decisions are made and Z^{er}_t is realized, the coordinator chooses a pair of *prescriptions*, $\Lambda_t := (\Lambda^1_t, \Lambda^2_t)$, where Λ^i_t is a mapping from X^i_t to U^i_t (more precisely, Λ^i_t maps $\mathcal{X}^i$ to $\mathcal{U}^i$). Agent i then generates its control action by evaluating the function Λ^i_t on its current local state: $U^i_t = \Lambda^i_t(X^i_t)$. The coordinator chooses its prescriptions based on the common information. Thus

$$\Gamma^1_t = d^1_t(C_t),\ \Gamma^2_t = d^2_t(C_t)$$

$$\Lambda^1_t = d^1_{t+}(C_{t+}),\ \Lambda^2_t = d^2_{t+}(C_{t+}) \tag{10}$$

where $d^1_t, d^2_t, d^1_{t+}, d^2_{t+}$ are referred to as the *coordinator's communication and control strategy for the two agents at time t*. The collection of functions $(d^1_1, d^2_1, d^1_{1+}, \ldots, d^1_{T+}, d^2_{T+})$ is called the coordinator's strategy. The coordinator's strategy optimization problem is to find a coordination strategy to minimize the expected total cost given by (5). The following lemma shows the equivalence of the coordinator's strategy optimization problem and the original strategy optimization problem for the agents.

Lemma 2: Suppose that $(d^{1*}_1, d^{2*}_1, \ldots, d^{1*}_{T+}, d^{2*}_{T+})$ is an optimal strategy for the coordinator. Then, optimal communication and control strategies for the agents in the original problem can be obtained as follows: for $i = 1, 2$,

$$\bar{f}^{i*}_t(X^i_t, C_t) = \Gamma^i_t(X^i_t) \text{ where} \Gamma^i_t = d^{i*}_t(C_t) \tag{11}$$

$$\bar{g}^{i*}_t(X^i_t, C_{t+}) = \Lambda^i_t(X^i_t) \text{ where} \Lambda^i_t = d^{i*}_{t+}(C_{t+}). \tag{12}$$

Proof: The lemma is a direct consequence of the results in [10]. ■

Lemma 2 implies that the agents' strategy optimization problem can be solved by solving the coordinator's strategy optimization problem. The advantage of the coordinator's problem is that it is a sequential decision-making problem with the coordinator as the only decision-maker. (Note that once the coordinator makes its decisions about which prescription to use, the agents act as mere evaluators and not as independent decision-makers.)

Coordinator's belief state: As shown in [10], the coordinator's problem can be viewed as a POMDP. Therefore, the coordinator's belief state can serve as the sufficient statistic for selecting prescriptions. Before communication at time t, the coordinator's belief is given as

$$\Pi_t(x^1, x^2) = \mathrm{P}(X^1_t = x^1, X^2_t = x^2|C_t, \Gamma_{1:(t-1)}, \Lambda_{1:(t-1)}). \tag{13}$$

After the communication decisions are made and Z^{er}_t is realized, the coordinator's belief is given as

$$\Pi_{t+}(x^1, x^2) = \mathrm{P}(X^1_t = x^1, X^2_t = x^2|C_{t+}, \Gamma_{1:t}, \Lambda_{1:(t-1)}). \tag{14}$$

Because of the conditional independence property identified in Lemma 1, the coordinator's beliefs can be factorized into beliefs on each agent's state, i.e.,

$$\Pi_t(x^1, x^2) = \Pi^1_t(x^1)\Pi^2_t(x^2) \tag{15}$$

$$\Pi_{t+}(x^1, x^2) = \Pi^1_{t+}(x^1)\Pi^2_{t+}(x^2) \tag{16}$$

where, for $i = 1, 2$, Π^i_t is the marginal belief on X^i_t obtained from 13 and Π^i_t+ is the marginal belief on X^i_t obtained from 14. The coordinator can update its beliefs on the agents' states in a sequential manner as described in the following lemma.

Lemma 3: For $i = 1, 2$, Π^i_1 is the prior belief $(P_{X^i_1})$ on the initial state X^i_1 and for each $t \geq 1$

$$\Pi^i_{t+} = \eta^i_t(\Pi^i_t, \Gamma^i_t, Z^{er}_t, M_t) \tag{17}$$

$$\Pi^i_{t+1} = \beta^i_t(\Pi^i_{t+}, \Lambda^i_t) \tag{18}$$

where η^i_t and β^i_t are fixed functions derived from the system model. (We will use $\beta_t(\Pi^{1,2}_{t+}, \Lambda^{1,2}_t)$ to denote the pair $\beta^1_t(\Pi^1_{t+}, \Lambda^1_t), \beta^2_t(\Pi^2_{t+}, \Lambda^2_t)$. Similar notation will be used for the pair $\eta^1_t(\cdot), \eta^2_t(\cdot).$)

Proof: The proof follows from Bayes' rule and the system model. The exact form of the belief update functions η^i_t, β^i_t is given in Appendix C. ■

Finally, we note that given the coordinator beliefs Π^1_t, Π^2_t and its prescriptions Γ^1_t, Γ^2_t at time t, the joint probability that $Z^{er}_t = \phi$ and $M_t = m_t$ is given as

$$P(Z^{er}_t = \phi, M_t = m_t|\Pi^{1,2}_t, \Gamma^{1,2}_t) = \begin{cases} \sum_{x^1,x^2} \mathbb{1}_{\{\Gamma^1_t(x^1)=0\}} \mathbb{1}_{\{\Gamma^2_t(x^2)=0\}} \Pi^1_t(x^1)\Pi^2_t(x^2) \text{ if } m_t = (0,0) \\ \sum_{x^1,x^2} p_e \mathbb{1}_{\{\Gamma^1_t(x^1)=m^1_t\}} \mathbb{1}_{\{\Gamma^2_t(x^2)=m^2_t\}} \Pi^1_t(x^1)\Pi^2_t(x^2) \text{ otherwise.} \end{cases} \tag{19}$$

Similarly, the probability that $Z^{er}_t = (x^1, x^2)$ is given as

$$P(Z^{er}_t = (x^1, x^2)|\Pi^{1,2}_t, \Gamma^{1,2}_t) = (1 - p_e)\left[\max(\Gamma^1_t(x^1), \Gamma^2_t(x^2))\right] \Pi^1_t(x^1)\Pi^2_t(x^2). \tag{20}$$

Authorized licensed use limited to: University of Southern California. Downloaded on October 16,2024 at 18:46:09 UTC from IEEE Xplore. Restrictions apply.

Algorithm 1: Strategies f^{i*}, g^{i*} for Agent i in the Team.

Input: $\Xi_t(\cdot), \Xi_{t^+}(\cdot)$ obtained from DP for all t
for $t = 1$ **to** T **do**
Before communication:
Current information: C_t, X_t^i {where $C_t = C_{(t-1)^+}$}
Update CIB $\Pi_t = \beta_{t-1}(\Pi_{(t-1)^+}, \Xi^1_{(t-1)^+}(\Pi_{t-1^+}))$ {If $t = 1$, Initialize CIB Π_t using C_1}
Get prescription $\Gamma_t = (\Gamma_t^1, \Gamma_t^2) = \Xi_t(\Pi_t)$
Select communication action $M_t^i = \Gamma_t^i(X_t^i)$
After communication decisions are made:
Current information: $C_{t^+}, X^i_{t^+}$ {where $C_{t^+} = \{C_t, Z_t^{er}, M_t\}$}
Update CIB $\Pi_{t^+} = \eta_t(\Pi_t, \Xi_t^1(\Pi_t), Z_t^{er}, M_t)$
Get prescription $\Lambda_t = (\Lambda_t^1, \Lambda_t^2) = \Xi_{t^+}(\Pi_{t^+})$
Select control action $U_t^i = \Lambda_t^i(X_t^i)$
end for

Coordinator's dynamic program: Using Lemma 3 and the probabilities given in (19) and (20), we can write a dynamic program for the coordinator's POMDP problem. In the following theorem, π^i denotes a general probability distribution on $\mathcal{X}^i$ and δ_{x^i} denotes a delta distribution centered at x^i.

Theorem 1: The value functions for the coordinator's dynamic program are as follows: For all beliefs π^1, π^2, $V_{T+1}(\pi^1, \pi^2) := 0$ and for $t = T, \ldots, 2, 1$, $V_{t^+}(\pi^1, \pi^2)$

$$:= \min_{\lambda^1, \lambda^2} \Bigg[\left(\sum_{x^{1,2}} c_t\left(x^{1,2}, \lambda^1(x^1), \lambda^2(x^2)\right) \pi^1(x^1)\pi^2(x^2) \right) + V_{t+1}(\beta_t(\pi^{1,2}, \lambda^{1,2})) \Bigg] \tag{21}$$

where β_t is as described in Lemma 3 and $V_t(\pi^1, \pi^2)$

$$:= \min_{\gamma^1, \gamma^2} \Bigg[\sum_{x^{1,2}} \rho(x^{1,2}) \max(\gamma^1(x^1), \gamma^2(x^2))\pi^1(x^1)\pi^2(x^2) + \sum_m P(Z_t^{er} = \phi, M_t = m | \pi^{1,2}, \gamma^{1,2}) V_{t^+}(\eta_t(\pi^{1,2}, \gamma^{1,2}, \phi, m)) + \sum_{\tilde{x}^{1,2}} P(Z_t^{er} = \tilde{x}^{1,2} | \pi^{1,2}, \gamma^{1,2}) V_{t^+}(\delta_{\tilde{x}^1}, \delta_{\tilde{x}^2}) \Bigg] \tag{22}$$

where η_t is as described in Lemma 3 and $P(Z_t^{er} = \phi, M_t = m | \pi^{1,2}, \gamma^{1,2})$, $P(Z_t^{er} = \tilde{x}^{1,2} | \pi^{1,2}, \gamma^{1,2})$ are as described in (19) and (20). The coordinator's optimal strategy is to pick the minimizing prescription pairs for each time and each (π^1, π^2).

Proof: Since the coordinator's problem is a POMDP, it has a corresponding dynamic program. The value functions in the theorem can be obtained by simple manipulations of the POMDP dynamic program for the coordinator. ■

Let $\Xi_t(\pi_t)$ (resp. $\Xi_{t^+}(\pi_{t^+})$) be a minimizer of the value function in (21) (resp. (22)). Using $\Xi_t(\pi_t), \Xi_{t^+}(\pi_{t^+})$ obtained from the dynamic program in Theorem 1, we can construct a strategy pair f^*, g^* as described in Algorithm 1. This strategy pair is an optimal strategy pair for the agents in the original problem.

The common-information-based dynamic program provided by (21) and (22) can be solved using off-the-shelf POMDP solvers [19]. Even though the common information beliefs (π_t^1, π_t^2) lie in a finite-dimensional continuous space, POMDP solvers exploit the piecewise-linearity and convexity of the value functions and provide tractable solutions to the dynamic program. We emphasize that the common-information-based dynamic program is solved offline. Therefore, the solution Ξ_t to this dynamic program is known to both agents before the agents start operating in their environment. During their operation, this solution is used in a decentralized manner by the agents to select their actions, as described in Algorithm 1.

The complexity of common-information-based dynamic programs (like the one in Theorem 1) depends largely on the size of the private information space. If we directly adopt the result in [10] without any private information reduction, the state space in the coordinator's POMDP would involve the history of local states. This space grows exponentially with time. The action space in the coordinator's POMDP would be the space of mappings from the history of local states to actions/decisions, which would grow doubly exponentially in time. These incredibly large state and action space sizes would make it impossible to use any POMDP solver to solve these problems. Our main contribution is to show that this computation complexity can be reduced substantially because a large part of the private information can be ignored without the loss of optimality (see Proposition 1). This result, in combination with ideas from the work in [10], allows us to obtain Theorem 1.

Remark 2: If the transition and cost functions are time invariant, we can also consider an infinite-horizon discounted cost analog of the problem formulation in this article. The previous results can be extended to this discounted setting in a straightforward manner using the approach in [10].

V. Extensions

A. Packet-Drop Channel With State

In our formulation in Section II, the quality of the communication channel between the agents did not change with time. In this section, we consider an extension where the packet-drop probability evolves over time as an uncontrolled Markov process. Let $E_t \in \mathcal{E}$ denote the channel state at time t where $\mathcal{E}$ is a finite set of channel states. The process E_t evolves as $E_{t+1} = l_t(E_t, W_t^e)$, where the random variables $\{W_t^e\}_{t \geq 1}$ are mutually independent and also independent of all the other primitive random variables. The packet-drop probability of the channel at time t, denoted by p_{e_t}, is a function of the channel state E_t, i.e., $p_{e_t} = \varphi_t(E_t)$. Furthermore, the communication cost also depends on the channel state and is given by $\rho(X_t, E_t)$.

The channel state is known to both agents. The information available to agent i at times t (before communication) and t^+ (after communication) is, thus, given by

$$I_t^i = \{X_{1:t}^i, U_{1:t-1}^i, Z_{1:t-1}^{er}, M_{1:t-1}^{1,2}, E_{1:t}\} \tag{23}$$

$$I_{t^+}^i = \{I_t^i, Z_t^{er}, M_t^{1,2}\}. \tag{24}$$

Our goal is to find communication and control strategies for the agents in the earlier setup. With some minor modifications, we can use the common-information-based methodology of Section IV to solve this problem.

Given the channel state E_t, the coordinator beliefs Π_t^1, Π_t^2 and its prescriptions Γ_t^1, Γ_t^2 at time t, the joint probability that $Z_t^{er} = \phi$ and $M_t = m_t$ is given as

$$P(Z_t^{er} = \phi, M_t = m_t | \Pi_t^{1,2}, \Gamma_t^{1,2}, E_t) = \begin{cases} \sum_{x^{1,2}} \mathbb{1}_{\{\Gamma_t^1(x^1)=0\}} \mathbb{1}_{\{\Gamma_t^2(x^2)=0\}} \Pi_t^1(x^1)\Pi_t^2(x^2), & \text{if } m_t = (0,0) \\ \sum_{x^{1,2}} \varphi_t(E_t) \mathbb{1}_{\{\Gamma_t^1(x^1)=m_t^1\}} \mathbb{1}_{\{\Gamma_t^2(x^2)=m_t^2\}} \Pi_t^1(x^1)\Pi_t^2(x^2), & \\ & \text{otherwise.} \end{cases} \tag{25}$$

Similarly, the probability that $Z_t^{er} = (x^1, x^2)$ is given as

$$P(Z_t^{er} = (x^1, x^2) | \Pi_t^{1,2}, \Gamma_t^{1,2}, E_t) = (1 - \varphi_t(E_t)) \left[\max(\Gamma_t^1(x^1), \Gamma_t^2(x^2))\right] \Pi_t^1(x^1)\Pi_t^2(x^2). \tag{26}$$

The following theorem describes the modified dynamic program for the coordinator. The value functions and the coordinator's optimal strategy depend on the current channel state in addition to the coordinator's beliefs.

Authorized licensed use limited to: University of Southern California. Downloaded on October 16,2024 at 18:46:09 UTC from IEEE Xplore. Restrictions apply.

Theorem 2: The value functions for the coordinator's dynamic program are as follows: For all beliefs π^1, π^2, and all $e \in \mathcal{E}$, $V_{T+1}(\pi^1, \pi^2, e) := 0$ and for $t = T, \dots, 2, 1$, $V_{t^+}(\pi^1, \pi^2, e)$

$$:= \min_{\lambda^1, \lambda^2} \Bigg[\sum_{x^{1,2}} c_t \left(x^{1,2}, \lambda^1(x^1), \lambda^2(x^2)\right) \pi^1(x^1)\pi^2(x^2)$$

$$+ \mathbb{E}[V_{t+1}(\beta_t(\pi^{1,2}, \lambda^{1,2}), E_{t+1}) \mid E_t = e] \Bigg] \tag{27}$$

where β_t is as described in Lemma 3, and $V_t(\pi^1, \pi^2, e)$

$$:= \min_{\gamma^1, \gamma^2} \Bigg[\sum_{x^{1,2}} \rho(x^{1,2}, e) \max(\gamma^1(x^1), \gamma^2(x^2)) \pi^1(x^1)\pi^2(x^2)$$

$$+ \sum_{m} P(Z_t^{er} = \phi, M_t = m | \pi^{1,2}, \gamma^{1,2}, e) V_{t^+}(\eta_t(\pi^{1,2}, \gamma^{1,2}, \phi, m), e)$$

$$+ \sum_{\tilde{x}^{1,2}} P(Z_t^{er} = \tilde{x}^{1,2} | \pi^{1,2}, \gamma^{1,2}, e) V_{t^+}(\delta_{\tilde{x}^1}, \delta_{\tilde{x}^2}, e) \Bigg] \tag{28}$$

where η_t is as described in Lemma 3 and $P(Z_t^{er} = \phi, M_t = m | \pi^{1,2}, \gamma^{1,2}, e)$, $P(Z_t^{er} = \tilde{x}^{1,2} | \pi^{1,2}, \gamma^{1,2}, e)$ are as described in (25)–(26). The coordinator's optimal strategy is to pick the minimizing prescription pairs for each time and each (π^1, π^2, e).

B. Agents With Communication Constraints

In this section, we consider an extension of the problem formulated in Section II where we incorporate some constraints on the communication between agents. The underlying system model, information structure, and the total expected cost are the same as in Section II. But now agents have constraints on when and how frequently they can communicate. Specifically, we consider the following three constraints.

1) Minimum time between successive communication attempts (i.e., times at which $M_t^{or} = 1$) must be at least $s_{\min}$ (where $s_{\min} \geq 0$).
2) Maximum time between successive communication attempts cannot exceed $s_{\max}$ (where $s_{\max} \geq s_{\min}$).
3) The total number of communication attempts over the time horizon T cannot exceed N.

The strategy optimization problem is to find communication and control strategies for the agents that minimize the expected cost in (5) while ensuring that the above three constraints are satisfied. We assume that there is at least one choice of agents' strategies for which the constraints are satisfied (i.e., the constrained problem is feasible). Note that our framework allows for some of the previous three constraints to be absent (e.g., setting $s_{\min} = 0$ effectively removes the first constraint; setting $N = T$ effectively removes the third constraint).

We can follow the methodology of Section IV for the constrained problem as well. The key difference is that in addition to the coordinator's beliefs on the agents' states, we will also need to keep track of 1) the time since the most recent communication attempt (denoted by S_t^a), and 2) the total number of communication attempts so far (denoted by S_t^b). The variables S_t^a, S_t^b are used by the coordinator to ensure that the prescriptions it selects will not result in constraint violations. For example, if $S_t^a < s_{\min}$, the coordinator can only select the communication prescriptions that map $\mathcal{X}^i$ to 0 for each i since this ensures that the first constraint will be satisfied. Similarly, if $S_t^a = s_{\max}$, then the coordinator must select a pair of communication prescriptions that ensure that communication happens at the current time. The following theorem describes the modified dynamic program for the coordinator in the constrained formulation.

Theorem 3: The value functions for the coordinator's dynamic program are as follows: For all beliefs π^1, π^2, and all non-negative integers s^a, s^b, $V_{T+1}(\pi^1, \pi^2, s^a, s^b) := 0$ and for $t = T, \dots, 2, 1$, $V_{t^+}(\pi^1, \pi^2, s^a, s^b)$

$$:= \min_{\lambda^1, \lambda^2} \Bigg[\sum_{x^{1,2}} c_t \left(x^{1,2}, \lambda^1(x^1), \lambda^2(x^2)\right) \pi^1(x^1)\pi^2(x^2)$$

$$+ V_{t+1}(\beta_t(\pi^{1,2}, \lambda^{1,2}), s^a, s^b) \Bigg] \tag{29}$$

where β_t is as described in Lemma 3; and if $s_{\min} \leq s^a < s_{\max}$ and $s^b < N$, $V_t(\pi^1, \pi^2, s^a, s^b)$

$$:= \min_{\gamma^1, \gamma^2} \Bigg[\sum_{x^{1,2}} \rho(x^{1,2}) \max(\gamma^1(x^1), \gamma^2(x^2)) \pi^1(x^1)\pi^2(x^2)$$

$$+ P(Z_t^{er} = \phi, M_t = (0,0) | \pi^{1,2}, \gamma^{1,2})$$

$$\times V_{t^+}(\eta_t(\pi^{1,2}, \gamma^{1,2}, \phi, m = (0,0)), s^a + 1, s^b)$$

$$+ \sum_{m \neq (0,0)} P(Z_t^{er} = \phi, M_t \neq (0,0) | \pi^{1,2}, \gamma^{1,2})$$

$$\times V_{t^+}(\eta_t(\pi^{1,2}, \gamma^{1,2}, \phi, m \neq (0,0)), 0, s^b + 1)$$

$$+ \sum_{\tilde{x}^{1,2}} P(Z_t^{er} = \tilde{x}^{1,2} | \pi^{1,2}, \gamma^{1,2}) V_{t^+}(\delta_{\tilde{x}^1}, \delta_{\tilde{x}^2}, 0, s^b + 1) \Bigg] \tag{30}$$

where η_t is as described in Lemma 3 and $P(Z_t^{er} = \phi, M_t = m | \pi^{1,2}, \gamma^{1,2})$, $P(Z_t^{er} = \tilde{x}^{1,2} | \pi^{1,2}, \gamma^{1,2})$ are as described by (19)–(20). If $s^b = N$ or if $s^a < s_{\min}$, then the minimization over γ^1, γ^2 in (30) is replaced by simply setting γ^1, γ^2 to be the prescriptions that map all states to 0. If $s^b < N$ and if $s^a = s_{\max}$, then the minimization over γ^1, γ^2 in (30) is replaced by simply setting γ^1, γ^2 to be the prescriptions that map all states to 1. The coordinator's optimal strategy is to pick the minimizing prescription pairs for each time and each (π^1, π^2, s^a, s^b).

TABLE I
TRANSITION PROBABILITIES $\mathrm{P}[X_{t+1}^i \mid X_t^i, U_t^i]$

	$U_t^i = \aleph$	$U_t^i = d_1$	$U_t^i = d_2$
$X_t^i = 0$	$(1\text{-}p_a^i, p_a^i)$	$(1\text{-}p_a^i, p_a^i)$	$(1\text{-}p_a^i, p_a^i)$
$X_t^i = 1$	$(0,1)$	$(p_{d_1}^i, 1\text{-}p_{d_1}^i)$	$(p_{d_2}^i, 1\text{-}p_{d_2}^i)$

VI. Illustrative Example

Problem setup: Consider a system where there are two entities that are susceptible to attacks. Each entity has an associated defender that can make decisions about whether and how to defend the entity. The defenders can take one of three possible actions: $\aleph$ (which denotes doing nothing), d_1, d_2. Thus, the defenders are the decision-making agents in this model. The state $X_t^i \in \{0, 1\}$ of agent i represents whether or not entity i is under attack at time t. We use 1 to denote the *attack* state and 0 to denote the *safe* (nonattack) state. If entity i is currently in the safe state, i.e., $X_t^i = 0$, then with probability p_a^i, the entity transitions to the attack state 1 (irrespective of the defender's action). When entity i is under attack, i.e., $X_t^i = 1$, if the corresponding defender chooses to do nothing, then the state does not change with probability 1, i.e., $X_{t+1}^i = X_t^i$. On the other hand, if the defender chooses to defend using defensive action d_k, where $k = 1, 2$, then the entity transitions to the safe state 0 with probability $p_{d_k}^i$. The transition probabilities are listed in a tabular form in Table I. If both entities are in the safe state, then the cost incurred by the system is 0. If at least one entity is under attack, then the cost incurred is 20. Furthermore, an additional cost of 100 (respectively 150) is incurred if both defenders choose to defend using d_1 (resp. d_2) at the same time (in any state). More explicitly, the cost at time t is given by

$$c_t(X_t, U_t) = \vartheta^{t-1} \Big[20 \mathbb{1}_{(X_t^1 = 1 \text{ or } X_t^2 = 1)} + 100 \mathbb{1}_{(U_t^1 = d^1)} \mathbb{1}_{(U_t^2 = d^1)}$$

$$+ 150 \mathbb{1}_{(U_t^1 = d^2)} \mathbb{1}_{(U_t^2 = d^2)} \Big]$$

Authorized licensed use limited to: University of Southern California. Downloaded on October 16,2024 at 18:46:09 UTC from IEEE Xplore. Restrictions apply.

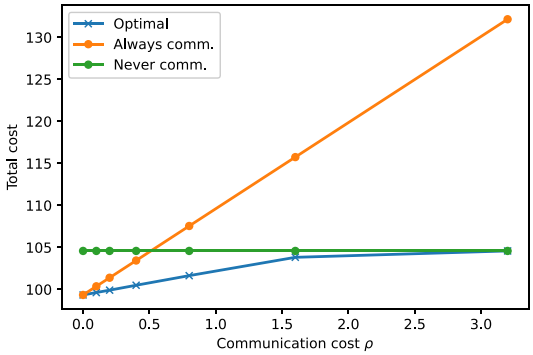

Fig. 1. Performance achieved by three strategies: 1) Jointly optimal communication and control strategies, 2) always communicate, and 3) never communicate. In this example, $p_a^1 = p_a^2 = 0.3$, $p_{d_1}^1 = p_{d_1}^2 = 0.6$, $p_{d_2}^1 = p_{d_2}^2 = 0.4$, and $\vartheta = 0.95$.

where $0 < \vartheta < 1$ is a discount factor. We assume that the pack-drop probability $p_e = 0$ and that the communication cost is a constant ρ.

When an entity is under attack, the associated defender needs to defend the system at some point of time (if not immediately). Otherwise, the system will remain in the attack state perpetually. However, a heavy cost is incurred if both agents defend using the same defensive action. Therefore, the agents must defend their respective entities in a coordinated manner. Communicating with each other can help agents coordinate effectively. On the other hand, communicating all the time can lead to a high communication cost. This tradeoff between communication and coordination can be balanced optimally using our approach discussed in Section IV.

Implementation: In our experiments, we consider an infinite horizon discounted cost version of the problem described earlier. Since the agents alternate between communication and control (see Section IV), the coordinator's POMDP as described is not time-invariant. To convert it into a time-invariant POMDP, we introduce an additional binary state variable X_t^c. This variable represents whether the agents currently are in the communication phase or the control phase. The variable X_t^c alternates between 0 and 1 in a deterministic manner. For agent i in the communication phase, action $\aleph$ is interpreted as the no communication decision ($M_t^i = 0$) and all other actions are interpreted as the communication decision ($M_t^i = 1$). With this transformation, we can use any infinite horizon POMDP solver to obtain approximately optimal strategies for our problem. In our experiments, we use the SARSOP solver [19] that is available in the Julia POMDPs framework [20]. The computational time did not exceed 1000 s, which was the specified time limit for the SARSOP POMDP solver.

Results: We consider three scenarios in our experiments: 1) the jointly (approximately) optimal communication and control strategies computed using the coordinator's POMDP; 2) the "never-communicate" communication strategy for agents along with control strategies that are optimized assuming no communication; and 3) the "always-communicate" communication strategy for agents along with control strategies that are optimizing assuming persistent communication. The total expected costs associated with these three strategies are shown in Fig. 1 for different choices of the c ommunication cost parameter ρ. The approximation error achieved using the SARSOP solver is at most 0.001.

Scaling up: A major challenge in solving the coordinator's POMDP is that the size of the prescription space is exponential in the size of the state space $\mathcal{X}^i$. Recall that in the coordinator's POMDP, the coordinator's action space is the space of prescription pairs. POMDP solvers need to optimize over the POMDPs action space (in our case prescription space) repeatedly as an intermediate step [19]. A naive approach for optimizing over prescription space is to enumerate every prescription pair and choose the one with the optimum value. This approach is commonly used when the action space in POMDPs is fairly small. In [7] and [21], an approach based on constraint optimization was proposed to tackle the computational complexity involved in the exhaustive enumeration of all prescriptions. It was noted in [7] and [21] that this approach works significantly better in practice. Our current implementation based on the Julia framework can be used only when the prescription space is small. In order to solve large-scale problems, one can modify the algorithm in [19] to incorporate the constraint optimization approach of [7] and [21].

VII. Conclusion

We considered a multiagent problem where agents can dynamically decide at each time step whether to share information with each other and incur the resulting communication cost. Our goal was to jointly design agents' communication and control strategies in order to optimize the tradeoff between communication costs and control objectives. We showed that agents can ignore a big part of their private information without compromising the system performance. We then provided a common-information-approach-based solution for the strategy optimization problem. Our approach relies on constructing a fictitious POMDP whose solution (obtained via a dynamic program) characterizes the optimal strategies for agents. We extended our solution to incorporate time-varying packet-drop channels and constraints on when and how frequently agents can communicate. A multiagent system in which a decentralized team of agents controls a stochastic system in the presence of an adversary is left for future work. One bottleneck we observed is the minimization over prescription space and we need more efficient ways to solve it. Using identical prescriptions for two agents is one simple way of minimizing the prescription space and will be explored in our future work.

Appendix

A. Proof of Lemma 1

We prove the lemma by induction. At $t = 1$, before communication decisions are made, (6) is trivially true since there is no common information at this point and the agents' initial states are independent.

Induction step: Suppose that (6) holds at time t. Then, we can show that (7) holds at time t^+. In order to do so, it suffices to show that the left-hand side of (7) can be factorized as follows:

$$\mathrm{P}(x_{1:t}, u_{1:t} | c_{t^+}) = \chi^1(x_{1:t}^1, u_{1:t}^1, c_{t^+}) \chi^2(x_{1:t}^2, u_{1:t}^2, c_{t^+}) \tag{31}$$

where χ^1 and χ^2 are some real-valued mappings with χ^i depending only on agent i's strategy. We now factorize the joint distribution ahead. Recall that $c_{t^+} = (c_{(t-1)^+}, z_t^{er}, m_t)$, $I_{t^+}^i = (x_{1:t}^i, u_{1:t-1}^i, z_{1:t}^{er}, m_{1:t})$, and $I_t^i = (x_{1:t}^i, u_{1:t-1}^i, z_{1:t-1}^{er}, m_{1:t-1})$. The left-hand side of (7) for t^+ can be written as

$$\begin{aligned}
&\frac{\mathrm{P}(x_{1:t}, u_{1:t}, z_t^{er}, m_t | c_{(t-1)^+})}{\mathrm{P}(z_t^{er}, m_t | c_{(t-1)^+})} \\
&= \mathrm{P}(u_t | x_{1:t}, u_{1:t-1}, c_{t^+}) \, \mathrm{P}(z_t^{er} | x_{1:t}, u_{1:t-1}, c_{(t-1)^+}, m_t) \\
&\quad \times \frac{\mathrm{P}(m_t | x_{1:t}, u_{1:t-1}, c_{(t-1)^+}) \, \mathrm{P}(x_{1:t}, u_{1:t-1} | c_{(t-1)^+})}{\mathrm{P}(z_t^{er}, m_t | c_{(t-1)^+})} \\
&= \left(\mathbb{1}_{(m_t^1 = f_t^1(I_t^1))} \mathbb{1}_{(u_t^1 = g_t^1(I_{t^+}^1))} \, \mathrm{P}(x_{1:t}^1, u_{1:t-1}^1 | c_{(t-1)^+}) \right) \\
&\quad \times \left(\mathbb{1}_{(m_t^2 = f_t^2(I_t^2))} \mathbb{1}_{(u_t^2 = g_t^2(I_{t^+}^2))} \, \mathrm{P}(x_{1:t}^2, u_{1:t-1}^2 | c_{(t-1)^+}) \right) \\
&\quad \times \frac{\mathrm{P}(z_t^{er} | x_{1:t}, u_{1:t-1}, c_{(t-1)^+}, m_t)}{\mathrm{P}(z_t^{er}, m_t | c_{(t-1)^+})}
\end{aligned} \tag{32}$$

Authorized licensed use limited to: University of Southern California. Downloaded on October 16,2024 at 18:46:09 UTC from IEEE Xplore. Restrictions apply.

where the last equality follows from the fact that $c_{(t-1)^+} = c_t$ and the induction hypothesis at time t. Furthermore, we have

$$\mathrm{P}(z_t^{er}|x_{1:t}, u_{1:t-1}, c_{(t-1)^+}, m_t) = \begin{cases} 1, & \text{if } m_t = (0,0), z_t^{er} = \phi \\ p_e, & \text{if } m_t \neq (0,0), z_t^{er} = \phi \\ (1-p_e)\mathbb{1}_{\{x_t = (\tilde{x}_t^1, \tilde{x}_t^2)\}}, & \text{if } m_t \neq (0,0), z_t^{er} = (\tilde{x}_t^1, \tilde{x}_t^2) \end{cases} \tag{33}$$

which can clearly be factorized. From (32) and (33), the joint distribution $\mathrm{P}(x_{1:t}, u_{1:t}|c_{t^+})$ can be factorized as in (31) and, thus, (7) holds at time t^+. Using this result, we now show that our induction hypothesis holds at time $t+1$. Recall that $c_{t+1} = c_{t^+}$. At time $t+1$, before communication decisions are made, the left-hand side of (6) can be written as

$$\begin{aligned} &\mathrm{P}(x_{t+1}|x_{1:t}, u_{1:t}, c_{t+1})\, \mathrm{P}(x_{1:t}, u_{1:t}|c_{t+1}) \\ &= \mathrm{P}(x_{t+1}^2|x_{t+1}^1, x_{1:t}, u_{1:t}, c_{t+1})\, \mathrm{P}(x_{t+1}^1|x_{1:t}, u_{1:t}, c_{t+1}) \\ &\quad \times \mathrm{P}(x_{1:t}, u_{1:t}|c_{t^+}) \\ &= \mathrm{P}(x_{t+1}^1|x_t^1, u_t^1)\, \mathrm{P}(x_{1:t}^1, u_{1:t}^1|c_{t^+})\, \mathrm{P}(x_{t+1}^2|x_t^2, u_t^2)\, \mathrm{P}(x_{1:t}^2, u_{1:t}^2|c_{t^+}) \end{aligned} \tag{34}$$

where the last equation follows from the state dynamics in (1) and (7) at time t^+. Using the factored form of $\mathrm{P}(x_{1:t+1}, u_{1:t}|c_{t+1})$ in (34), we can conclude that our induction hypothesis holds at time $t+1$. Therefore, by induction, we can conclude that (6) and (7) hold at all times.

B. Proof of Proposition 1

We will prove the result for agent i. Throughout this proof, we fix agent $-i's$ communication and control strategies to be f^{-i}, g^{-i} (where f^{-i}, g^{-i} are arbitrarily chosen). Define $R_t^i = (X_t^i, Z_{1:t-1}^{er}, M_{1:t-1}^{1,2})$ and $R_{t^+}^i = (X_t^i, Z_{1:t}^{er}, M_{1:t}^{1,2})$. Our proof will rely on the following two facts.

Fact 1: $\{R_1^i, R_{1^+}^i, R_2^i, R_{2^+}^i, \ldots R_T^i, R_{T^+}^i\}$ is a controlled Markov process for agent i. More precisely, for any strategy choice f^i, g^i of agent i

$$\begin{aligned} &\mathrm{P}(R_{t^+}^i = \tilde{r}_{t^+}^i | R_{1:t}^i = r_{1:t}^i, M_{1:t}^i = m_{1:t}^i, U_{1:t-1}^i = u_{1:t-1}^i) \\ &\quad = \mathrm{P}(R_{t^+}^i = \tilde{r}_{t^+}^i | R_t^i = r_t^i, M_t^i = m_t^i) \end{aligned} \tag{35}$$

$$\begin{aligned} &\mathrm{P}(R_{t+1}^i = \tilde{r}_{t+1}^i | R_{1:t^+}^i = r_{1:t^+}^i, M_{1:t}^i = m_{1:t}^i, U_{1:t}^i = u_{1:t}^i) \\ &\quad = \mathrm{P}(R_{t+1}^i = \tilde{r}_{t+1}^i | R_{t^+}^i = r_{t^+}^i, U_t^i = u_t^i) \end{aligned} \tag{36}$$

where the probabilities on the right hand side of (35) and (36) do not depend on f^i, g^i.

Fact 2: The costs at time t satisfy

$$\mathbb{E}[\rho(X_t^i, X_t^{-i})\mathbb{1}_{(M_t^{cr}=1)}|r_{1:t}^i, m_{1:t}^i, u_{1:t-1}^i] = \kappa_t^i(r_t^i, m_t^i) \tag{37}$$

$$\mathbb{E}[c_t(X_t, U_t)|r_{1:t^+}^i, m_{1:t}^i, u_{1:t}^i] = \kappa_{t^+}^i(r_{t^+}^i, u_t^i) \tag{38}$$

where the functions $\kappa_t^i, \kappa_{t^+}^i$ in (37) and (38) do not depend on f^i, g^i.

Suppose that Facts 1 and 2 are true. Then, the strategy optimization problem for agent i can be viewed as an MDP over $2T$ time steps (i.e., time steps $1, 1^+, 2, 2^+, \ldots, T, T^+$) with R_t^i and M_t^i as the state and action at time t; and $R_{t^+}^i$ and U_t^i as the state and action for time t^+.

Note that at time t, agent i observes R_t^i, selects M_t^i and the "state" transitions to $R_{t^+}^i$ according to Markovian dynamics (35). Similarly, at time t^+, agent i observes $R_{t^+}^i$, selects U_t^i and the "state" transitions to R_{t+1}^i according to Markovian dynamics (36). Furthermore, from agent i's perspective, the cost at time t depends on the state and action at t [i.e., R_t^i and M_t^i, see (37)] and the cost at time t^+ depends on the state and action at t^+ [i.e., $R_{t^+}^i$ and U_t^i, see (38)]. It then follows from standard MDP results that agent i can find an optimal strategy (given agent $-i$'s strategy) of the form:

$$M_t^i = \bar{f}_t^i(R_t^i) = \bar{f}_t^i(X_t^i, Z_{1:t-1}^{er}, M_{1:t-1}^{1,2})$$
$$U_t^i = \bar{g}_t^i(R_{t^+}) = \bar{g}_t^i(X_t^i, Z_{1:t}^{er}, M_{1:t}^{1,2})$$

which establishes the result of the proposition (recall that $C_t = (Z_{1:t-1}^{er}, M_{1:t-1}^{1,2})$ and $C_{t^+} = (Z_{1:t}^{er}, M_{1:t}^{1,2})$), we now prove Facts 1 and 2 stated before.

Proof of Fact 1: Let $\tilde{r}_{t^+}^i = (x_t^i, z_{1:t}^{er}, m_{1:t})$ and $r_{1:t}^i = (x_{1:t}^i, z_{1:t-1}^{er}, m_{1:t-1})$. Then, the left-hand side of (35) can be written as

$$\begin{aligned} &\mathrm{P}(R_{t^+}^i = (x_t^i, z_{1:t}^{er}, m_{1:t}) | (x_{1:t}^i, z_{1:t-1}^{er}, m_{1:t-1}), m_{1:t}^i, u_{1:t-1}^i) \\ &= \mathrm{P}(z_t^{er}|x_{1:t}^i, z_{1:t-1}^{er}, m_{1:t}, u_{1:t-1}^i) \\ &\qquad \times \mathrm{P}(M_t^{-i} = m_t^{-i}|x_{1:t}^i, z_{1:t-1}^{er}, m_{1:t-1}, m_t^i, u_{1:t-1}^i) \\ &= \mathrm{P}(z_t^{er}|x_{1:t}^i, z_{1:t-1}^{er}, m_{1:t}, u_{1:t-1}^i)\, \mathrm{P}(m_t^{-i}|z_{1:t-1}^{er}, m_{1:t-1}) \end{aligned} \tag{39}$$

where (39) follows from the conditional independence property of Lemma 1. We can further simplify the first term in (39) for different cases as follows: $\mathrm{P}(z_t^{er}|x_{1:t}^i, z_{1:t-1}^{er}, m_{1:t}, u_{1:t-1}^i)$

$$= \begin{cases} (1-p_e)\mathbb{1}_{(\tilde{x}_t^i = x_t^i)}\, \mathrm{P}(\tilde{x}_t^{-i}|z_{1:t-1}^{er}, m_{1:t}), & \text{if } z_t^{er} = \tilde{x}_t, m_t \neq (0,0) \\ p_e, & \text{if } z_t^{er} = \phi, m_t \neq (0,0) \\ 1, & \text{if } z_t^{er} = \phi, m_t = (0,0). \end{cases}$$

We note that in all cases above $x_{1:t-1}^i$ does not affect the probability. This, combined with (39), establishes (35). We further note that the probabilities in the three cases above and in the second term of (39) do not depend on agent i's strategy. Equation (36) is a direct consequence of the Markovian state dynamics of agent i.

Proof of Fact 2: We have $\mathrm{P}[\tilde{x}_t, \tilde{m}_t|r_{1:t}^i, m_{1:t}^i, u_{1:t-1}^i]$

$$\begin{aligned} &= \mathrm{P}[\tilde{x}_t, \tilde{m}_t|x_{1:t}^i, z_{1:t-1}^{er}, m_{1:t-1}, m_t^i, u_{1:t-1}^i] \\ &= \mathbb{1}_{(\tilde{x}_t^i = x_t^i)}\mathbb{1}_{(\tilde{m}_t^i = m_t^i)}\, \mathrm{P}[X_t^{-i} = \tilde{x}_t^{-i}, M_t^{-i} = \tilde{m}_t^{-i}|z_{1:t-1}^{er}, m_{1:t-1}] \end{aligned}$$

where the last equation follows from the conditional independence in Lemma 1. Therefore, the probability distribution of X_t, M_t conditioned on $R_{1:t}^i, M_{1:t}^i$ depends only on $(x_t^i, z_{1:t-1}^{er}, m_{1:t-1}, m_t^i) = (r_t^i, m_t^i)$. Also note that this conditional probability does not depend on agent i's strategy. Hence, the conditional expectation in (37) can be expressed as a function of r_t^i, m_t^i. To prove (38), it suffices to show that

$$\begin{aligned} &\mathrm{P}(x_t^{-i}, u_t^{-i}|(x_{1:t}^i, z_{1:t}^{er}, m_{1:t}), m_{1:t}^i, u_{1:t}^i) \\ &= \mathrm{P}(x_t^{-i}, u_t^{-i}|(x_t^i, z_{1:t}^{er}, m_{1:t}), u_t^i) \quad \text{from Lemma 1.} \end{aligned}$$

C. Proof of Lemma 3

Let $c_t := (z_{1:t-1}^{er}, m_{1:t-1}^{1,2})$ and $c_{t^+} := (z_{1:t}^{er}, m_{1:t}^{1,2})$ be realizations of C_t, C_{t^+}, respectively. Let $c_{t+1} = c_{t^+}$ be the corresponding realization of C_{t+1}. Let $\gamma_{1:t}, \lambda_{1:t}$ be the realizations of the coordinator's prescriptions $\Gamma_{1:t}, \Lambda_{1:t}$ up to time t. Let us assume that realizations $c_{t+1}, \gamma_{1:t}, \lambda_{1:t}$ have nonzero probability. Let π_t^i, $\pi_{t^+}^i$, and π_{t+1}^i be the corresponding realizations of the coordinator's beliefs Π_t^i, $\Pi_{t^+}^i$, and Π_{t+1}^i, respectively. There are two possible cases: i) $z_t^{er} = (\tilde{x}_t^1, \tilde{x}_t^2)$, and ii) $z_t^{er} = \phi$. Let us analyze these cases separately.

Authorized licensed use limited to: University of Southern California. Downloaded on October 16,2024 at 18:46:09 UTC from IEEE Xplore. Restrictions apply.

Case (i): When $z_t^{er} = (\tilde{x}_t^1, \tilde{x}_t^2)$ for some $(\tilde{x}_t^1, \tilde{x}_t^2) \in \mathcal{X}^1 \times \mathcal{X}^2$, at least one of the agents must have decided to communicate at time t *and* the communication must have been successful. As described in (2), $Z_t^{er} = X_t$ when successful communication occurs. Thus, we have $\pi_{t^+}^i(x_t^i) = \mathrm{P}(X_t^i = x_t^i | z_{1:t}^{er}, m_{1:t}^{1,2}, \gamma_{1:t}, \lambda_{1:t-1}) = \mathbb{1}_{(x_t^i = \tilde{x}_t^i)}$.

Case (ii): In this case, $z_t^{er} = \phi$. Let $m_t := (m_t^1, m_t^2)$ and

$$q(m_t) := \mathrm{P}[Z_t^{er} = \phi \mid M_t = m_t] = \begin{cases} 1, & \text{if } m_t = (0,0) \\ p_e, & \text{otherwise.} \end{cases}$$

Using Bayes' rule, we can write $\pi_{t^+}^i(x_{t^+}^i)$ as

$$= \frac{\mathrm{P}(X_{t^+}^i = x_{t^+}^i, Z_t^{er} = \phi, M_t = m_t | z_{1:t-1}^{er}, m_{1:t-1}^{1,2}, \gamma_{1:t}, \lambda_{1:t-1})}{\mathrm{P}(Z_t^{er} = \phi, M_t = m_t | z_{1:t-1}^{er}, m_{1:t-1}^{1,2}, \gamma_{1:t}, \lambda_{1:t-1})}$$

$$= \frac{q(m_t)\,\mathrm{P}(M_t = m_t | x_t^i, c_t, \gamma_{1:t}, \lambda_{1:t-1})\,\mathrm{P}(x_t^i | c_t, \gamma_{1:t}, \lambda_{1:t-1})}{\sum_{\hat{x}_t^i} q(m_t)\,\mathrm{P}(M_t = m_t | \hat{x}_t^i, c_t, \gamma_{1:t}, \lambda_{1:t-1})\,\mathrm{P}(\hat{x}_t^i | c_t, \gamma_{1:t}, \lambda_{1:t-1})}$$

$$\overset{(a)}{=} \frac{\mathbb{1}_{(\gamma_t^i(x_t^i) = m_t^i)} \pi_t^i(x_t^i)}{\sum_{\hat{x}_t^i} \mathbb{1}_{(\gamma_t^i(\hat{x}_t^i) = m_t^i)} \pi_t^i(\hat{x}_t^i)}. \tag{40}$$

In (a), we use the fact that $\mathrm{P}(m_t | x_t^i, c_t, \gamma_{1:t}, \lambda_{1:t-1}) = P(m_t^{-i} | c_t, \gamma_{1:t}, \lambda_{1:t-1}) \mathbb{1}_{(\gamma_t^i(x_t^i) = m_t^i)}$. Hence, we can update the coordinator's belief $\pi_{t^+}^i(x_t^i)$ using π_t^i, γ_t^i, z_t^{er} and m_t as

$$\begin{cases} \dfrac{\mathbb{1}_{(\gamma_t^i(x_t^i) = m_t^i)} \pi_t^i(x_t^i)}{\sum_{\hat{x}_t^i} \mathbb{1}_{(\gamma_t^i(\hat{x}_t^i) = m_t^i)} \pi_t^i(\hat{x}_t^i)}, & \text{if} z_t^{er} = \phi \\ \mathbb{1}_{(x_t^i = \tilde{x}_t^i)}, & \text{if} z_t^{er} = (\tilde{x}_t^1, \tilde{x}_t^2). \end{cases} \tag{41}$$

We denote the update rule described earlier with η_t^i, i.e. $\pi_{t^+}^i = \eta_t^i(\pi_t^i, \gamma_t^i, z_t^{er}, m_t)$. Furthermore, using the law of total probability, we can write $\pi_{t+1}^i(x_{t+1}^i)$ as

$$= \sum_{x_t^i} \sum_{u_t^i} \Big[\mathrm{P}(x_{t+1}^i | x_t^i, u_t^i, z_{1:t}^{er}, m_{1:t}^{1,2}, \gamma_{1:t}, \lambda_{1:t})$$

$$\times \mathrm{P}(u_t^i | x_t^i, z_{1:t}^{er}, m_{1:t}^{1,2}, \gamma_{1:t}, \lambda_{1:t})\,\mathrm{P}(x_t^i | z_{1:t}^{er}, m_{1:t}^{1,2}, \gamma_{1:t}, \lambda_{1:t}) \Big]$$

$$= \sum_{x_t^i} \sum_{u_t^i} \mathrm{P}(x_{t+1}^i | x_t^i, u_t^i) \mathbb{1}_{(u_t^i = \lambda_t^i(x_t^i))} \pi_{t^+}^i(x_t^i). \tag{42}$$

We denote the update rule described earlier with β_t^i, i.e., $\pi_{t+1}^i = \beta_t^i(\pi_{t^+}^i, \lambda_t^i)$.

REFERENCES

[1] D. S. Bernstein, R. Givan, N. Immerman, and S. Zilberstein, "The complexity of decentralized control of Markov decision processes," *Math. Operations Res.*, vol. 27, no. 4, pp. 819–840, 2002.

[2] D. Szer, F. Charpillet, and S. Zilberstein, "MAA*: A heuristic search algorithm for solving decentralized POMDPs," in *Proc. 21st Conf. Uncertainty Artif. Intell.*, 2005, pp. 576–583.

[3] J. S. Dibangoye, C. Amato, O. Buffet, and F. Charpillet, "Optimally solving Dec-POMDPs as continuous-state MDPs," *J. Artif. Intell. Res.*, vol. 55, pp. 443–497, 2016.

[4] T. Rashid, M. Samvelyan, C. Schroeder, G. Farquhar, J. Foerster, and S. Whiteson, "QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 4295–4304.

[5] H. Hu and J. N. Foerster, "Simplified action decoder for deep multi-agent reinforcement learning," in *Proc. Int. Conf. Learn. Representations*, 2019.

[6] R. Becker, S. Zilberstein, V. Lesser, and C. V. Goldman, "Solving transition independent decentralized Markov decision processes," *J. Artif. Intell. Res.*, vol. 22, pp. 423–455, 2004.

[7] J. S. Dibangoye, C. Amato, and A. Doniec, "Scaling up decentralized MDPs through heuristic search," in *Proc. 28th Conf. Uncertainty Artif. Intell.*, 2012, pp. 217–226.

[8] Y. Xie, J. Dibangoye, and O. Buffet, "Optimally solving two-agent decentralized POMDPs under one-sided information sharing," in *Int. Conf. Mach. Learn.*, 2020, pp. 10473–10482.

[9] A. Nayyar, A. Mahajan, and D. Teneketzis, "Optimal control strategies in delayed sharing information structures," *IEEE Trans. Autom. Control*, vol. 56, no. 7, pp. 1606–1620, Jul. 2011.

[10] A. Nayyar, A. Mahajan, and D. Teneketzis, "Decentralized stochastic control with partial history sharing: A common information approach," *IEEE Trans. Autom. Control*, vol. 58, no. 7, pp. 1644–1658, Jul. 2013.

[11] A. Mahajan, "Optimal decentralized control of coupled subsystems with control sharing," *IEEE Trans. Autom. Control*, vol. 58, no. 9, pp. 2377–2382, Sep. 2013.

[12] J. Foerster et al., "Bayesian action decoder for deep multi-agent reinforcement learning," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 1942–1951.

[13] S. Sukhbaatar, A. Szlam, and R. Fergus, "Learning multiagent communication with backpropagation," in *Proc. 30th Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 2252–2260.

[14] J. N. Foerster, Y. M. Assael, N. de Freitas, and S. Whiteson, "Learning to communicate with deep multi-agent reinforcement learning," in *Proc. 30th Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 2145–2153.

[15] K. Cao, A. Lazaridou, M. Lanctot, J. Z. Leibo, K. Tuyls, and S. Clark, "Emergent communication through negotiation," in *Proc. Int. Conf. Learn. Representations*, 2018.

[16] A. Lazaridou and M. Baroni, "Emergent multi-agent communication in the deep learning era," 2020, *arXiv:2006.02419*.

[17] R. Lowe, Y. I. Wu, A. Tamar, J. Harb, O. P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6382–6393.

[18] J. Foerster, G. Farquhar, T. Afouras, N. Nardelli, and S. Whiteson, "Counterfactual multi-agent policy gradients," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 2974–2982.

[19] H. Kurniawati, D. Hsu, and W. S. Lee, "SARSOP: Efficient point-based POMDP planning by approximating optimally reachable belief spaces," in *Robotics: Science and Systems*, Princeton, NJ, USA: Citeseer, 2008.

[20] 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," *J. Mach. Learn. Res.*, vol. 18, no. 26, pp. 1–5, 2017. [Online]. Available: http://jmlr.org/papers/v18/16-300.html

[21] J. S. Dibangoye, C. Amato, A. Doniec, and F. Charpillet, "Producing efficient error-bounded solutions for transition independent decentralized MDPs," in *Proc. Int. Conf. Auton. Agents Multi-Agent Syst.*, 2013, pp. 539–546.

Authorized licensed use limited to: University of Southern California. Downloaded on October 16,2024 at 18:46:09 UTC from IEEE Xplore. Restrictions apply.