2022 American Control Conference (ACC)
Atlanta, USA, June 8-10, 2022

Decentralized Control of Two Agents with Nested Accessible Information

Aditya Dave, *Student Member, IEEE,* Nishanth Venkatesh, *Student Member, IEEE,*
Andreas A. Malikopoulos, *Senior Member, IEEE*

Abstract— **In this paper, we investigate a decentralized stochastic control problem with two agents, where a part of the memory of the second agent is also available to the first agent at each instance of time. We derive a structural form for optimal control strategies which allows us to restrict their domain to a set which does not grow in size with time. We also present a dynamic programming (DP) decomposition which can utilize our results to derive optimal strategies for arbitrarily long time horizons. Since obtaining optimal control strategies by solving this DP decomposition is computationally intensive, we present potential resolutions in the form of simplified strategies by imposing additional conditions on our model, and an approximation technique which can be used to implement our results with a bounded loss of optimality.**

I. Introduction

Decentralized stochastic control problems consist of cooperative agents who take actions over a time horizon to minimize a shared cost, with limited ability to communicate in real time. Typical decentralized systems include connected and automated vehicles [1] and social media platforms [2]. Generally, no single agent has both: (1) access to all information in the system and (2) the ability to assign all actions. Thus, such systems are characterized by their *information structure*, which describes the information available to each agent at each time. Various information structures, summarized in [3], are: *(1) Classical,* where all agents communicate and recall information perfectly [4]; *(2) Quasi-classical,* if agent 1 can affect the state of agent 2, and the information available to agent 1 is *also* available to agent 2 [5]; and *(3) Non-classical,* where agents can affect each others' states with incomplete information [6]–[13].

Non-classical systems suffer from doubly exponential growth in computations required to generate optimal control strategies with an increase in the planning horizon [14]. The common information approach [6] alleviates this problem for systems with partial history sharing among all agents. The main idea in this approach is to identify an *information state* which can be utilized, in place of the common information across all agents, to derive the optimal control strategies. For systems with partial history sharing, the information state and private information of each agent do not grow in size with time. Subsequently, we can use a dynamic programming (DP) decomposition to compute optimal control strategies for long time horizons. However, the computational tractability of this approach suffers if the private information of any agent grows with time. This phenomenon is commonly observed in systems with partially nested information [5], one-directional communication [7]–[10] and unreliable communication [15], [16]. For such systems, current methods focus on identifying specific dynamics and information structures which yield computationally tractable solutions [7], [8].

In this paper, we identify a general information structure, called *nested accessible information,* for decentralized systems with two agents, and show that even in the presence of noisy observations of the state, it yields control strategies which are functions of information states. At each instance of time, we consider that a subset of the information available to agent 2, called accessible information, is sequentially nested within the information available to agent 1. However, the information which is available to agent 1 and not available to agent 2 is allowed to grow in size with time. For example, this phenomenon occurs when agent 1 does not share their observations and actions with agent 2 but receives the actions and observations of agent 2 at each time. Other special cases of our information structure include teams of two agents with: (1) either instantaneous or delayed one-directional communication from agent 2 to agent 1 [10], [11]; (2) transmission of data from agent 1 to agent 2 using an unreliable communication channel, which was considered for systems with linear dynamics, quadratic costs and Gaussian noises in [16]; (3) real time communication from agent 1 to agent 2 [8].

Our main contribution in this paper is that we establish a structural form for optimal control strategies in systems with nested accessible information (Theorem 2). This structural form allows us to restrict their domain to a space which does not grow in size with time. In our exposition, we use a combination of the *person-by-person* [5] and *prescription* [12] approaches. While both approaches are well established, we combine them to yield results for optimal strategies which cannot be derived by an individual application of either of these approaches. Next, we present a DP decomposition which utilizes our results to obtain optimal control strategies (Section III-C). In general, solving this DP is computationally challenging. As a potential resolution, we show how our results can be simplified with an additional assumption of decoupled dynamics for the system (Section IV). Finally, we propose an approximate solution which can be used to improve the computational tractability of the DP even in the presence of coupled dynamics (Section V). While we

This research was supported by the Sociotechnical Systems Center (SSC) at the University of Delaware.

The authors are with the Department of Mechanical Engineering, University of Delaware, Newark, DE 19716 USA (email: adidave@udel.edu; nish@udel.edu; andreas@udel.edu).

978-1-6654-5197-0/$31.00 ©2022 AACC

restrict our attention to a team of two agents to simplify the exposition, our results can also be applied to systems with multiple agents in two nested subsystems, using a technique presented in Section III of [17].

The remainder of the paper is organized as follows. In Section II, we provide our problem formulation. In Section III, we analyze the problem and derive our main results. In Section IV, we present specialized results for systems with additional assumptions on the dynamics. In Section V, we present an approximation technique to implement our results. Finally, in Section VI, we present concluding remarks and discuss ongoing work.

II. Problem Formulation

We consider a team of two agents who take actions over $T \in \mathbb{N}$ discrete time steps. For each $t = 0, \dots, T$, the state of the team is denoted by the random variable X_t which takes values in a finite set $\mathcal{X}_t$. The action of an agent $k = 1, 2$ at time t is U_t^k, which takes values in a finite set $\mathcal{U}_t^k$. We denote the tuple (U_t^1, U_t^2) by $U_t^{1:2}$. Starting at the initial state X_0, the system evolves as

$$X_{t+1} = f_t\left(X_t, U_t^{1:2}, W_t\right), \quad t = 0, \dots, T-1, \qquad (1)$$

where W_t is an uncontrolled disturbance which takes values in a finite set $\mathcal{W}_t$. At each $t = 0, \dots, T$, each agent $k = 1, 2$ makes an observation $Y_t^k := h_t^k(X_t, V_t^k)$, which takes values in a finite set $\mathcal{Y}_t^k$. Here, V_t^k is a measurement noise which takes values in a finite set $\mathcal{V}_t^k$. The external disturbances $\{W_t : t = 0, \dots, T\}$, measurement noises $\{V_t^1, V_t^2 : t = 0, \dots, T\}$, and initial state X_0 are collectively called the *primitive random variables* of the team and their probability distributions are known a priori. We assume that each primitive random variable is independent of all other primitive random variables to ensure that the system's evolution is Markovian [4].

Definition 1. For all $t = 0, \dots, T$, the *memory* of an agent $k = 1, 2$ is a set of random variables $M_t^k \subseteq \{Y_{0:t}^{1:2}, U_{0:t-1}^{1:2}\}$, which takes values in a finite collection of sets $\mathcal{M}_t^k$ and satisfies *perfect recall*, i.e, $M_{t-1}^k \subseteq M_t^k$, with $M_{-1}^k := \emptyset$.

We partition the memory M_t^2 of agent 2 into two components, the *accessible information* A_t^2 and *private information* L_t^2, which are described next:

1) The accessible information is a subset of the memory of agent 2 which is also available to agent 1. For all $t = 0, \dots, T$, we define the accessible information as a set of random variables $A_t^2 \subseteq M_t^2$ which takes values in a finite collection of sets $\mathcal{A}_t^2$ and satisfies the properties: (1) accessibility to agent 1, i.e., $A_t^2 \subseteq M_t^1$, and (2) perfect recall, i.e., $A_{t-1}^2 \subseteq A_t^2$, with $A_{-1}^2 := \emptyset$.

2) The private information of agent 2 is a subset of their memory which is unavailable to agent 1. For all $t = 0, \dots, T$, we define the private information as the set of random variables $L_t^2 := M_t^2 \setminus A_t^2$ which takes values in a finite collection of sets $\mathcal{L}_t^2$. We impose the condition $L_t^2 \cap M_t^1 = \emptyset$ to specify that agent 1 can not access the private information of agent 2 at each t.

The second property of the accessible information of agent 2 motivates us to define the *new information* added to A_t^2, for all $t = 0, \dots, T$, as the set of random variables $Z_t^2 := A_t^2 \setminus A_{t-1}^2$ which takes values in a finite collection of sets $\mathcal{Z}_t^2$. Note that $Z_0^2 := A_0^2$. Analogously, for all $t = 0, \dots, T$, we define the new information added to the memory of agent 1 as the set of random variables $Z_t^1 := M_t^1 \setminus M_{t-1}^1$ which takes values in a finite collection of sets $\mathcal{Z}_t^1$, where $Z_0^1 := M_0^1$. In our information structure, we enforce that for all t, the new information of agent 2 must satisfy $Z_t^2 \subseteq L_t^2 \cup \{Y_t^{1:2}, U_{t-1}^{1:2}\}$. This ensures that $Z_t^2 \not\subset M_{t-1}^1$ and $Z_t^2 \subseteq Z_t^1$, i.e., Z_t^2 is not accessible to agent 1 prior to time t and becomes accessible to agent 1 at time t.

Remark 1. We call the shared set A_t^2 the *accessible information* of agent 2 instead of *common information* [6] to highlight the additional restriction imposed by the property $Z_t^2 \not\subset M_{t-1}^1$. The presence of this restriction allows us to specialize our results to systems where the information available to agent 1 but unavailable to agent 2, i.e., $M_t^1 \setminus A_t^2$, may grow in size with time. If we relax this restriction, the accessible information is equivalent to common information.

Remark 2. As an example of an information structure which satisfies $Z_t^2 \not\subset M_{t-1}^1$, consider one-directional communication from 2 to 1 with a delay of $d \in \mathbb{N}$ time steps. In such a system, $M_t^1 = \{Y_{0:t}^1, U_{0:t-1}^1, Y_{0:t-d}^2, U_{0:t-d}^2\}$ and $M_t^2 = \{Y_{0:t}^2, U_{0:t-1}^2\}$. Then, $A_t^2 = \{Y_{0:t-d}^2, U_{0:t-d}^2\}$, $L_t^2 = \{Y_{t-d+1:t}^2, U_{t-d+1:t-1}^2\}$, and the set $M_t^1 \setminus A_t^2 = \{Y_{0:t}^1, U_{0:t-1}^1\}$ grows in size with time. Recall that we have referenced other information structures which satisfy the conditions for nested accessible information in Section I.

For all $t = 0, \dots, T$, each agent $k = 1, 2$ uses a control law $g_t^k : \mathcal{M}_t^k \to \mathcal{U}_t^k$ to select their action

$$U_t^k = g_t^k(M_t^k), \qquad (2)$$

where $M_t^2 = \{L_t^2, A_t^2\}$. We define the control strategy of agent k as $\boldsymbol{g}^k := (g_t^k : t = 0, \dots, T)$ and the control strategy of the team as $\boldsymbol{g} := (\boldsymbol{g}^1, \boldsymbol{g}^2)$. The set of all feasible control strategies is $\mathcal{G}$. After each agent $k = 1, 2$ selects their action U_t^k at time t, the team incurs a cost $c_t(X_t, U_t^{1:2}) \in \mathbb{R}_{\geq 0}$. The performance criterion over the finite horizon T is

$$\mathcal{J}(\boldsymbol{g}) = \mathbb{E}^{\boldsymbol{g}}\left[\sum_{t=0}^{T} c_t\big(X_t, U_t^{1:2}\big)\right], \qquad (3)$$

where the expectation is with respect to the joint probability distribution on all random variables. Next, we state the optimization problem for the team.

Problem 1. The optimization problem for the team is $\inf_{\boldsymbol{g} \in \mathcal{G}} \mathcal{J}(\boldsymbol{g})$, given the distributions of the primitive random variables $\{X_0, W_{0:t}, V_{0:t}^{1:2}\}$, and the dynamics $\{c_t, f_t, h_t^{1:2} : t = 0, \dots, T\}$.

Problem 1 is guaranteed to have a solution because all variables take values in finite sets. Our goal is to derive a structural form for an optimal strategy $\boldsymbol{g}^* \in \mathcal{G}$ in Problem 1 which can be computed using a DP decomposition.

III. Analysis Using Prescriptions

A. Analysis for Agent 1

In this subsection, we derive a structural form for an optimal control strategy of agent 1. We first note that given a strategy $\boldsymbol{g}^2$, agent 1 cannot generate the action U_t^2 for each t because they cannot access the complete memory $M_t^2 = \{L_t^2, A_t^2\}$. However, they can access the component A_t^2. This motivates us to consider a two stage process for the generation of the action of agent 2: (1) agent 1 generates a prescription for agent 2 using only A_t^2, and (2) agent 2 computes U_t^2 using this prescription and their private information L_t^2.

Definition 2. For all $t = 0, \dots, T$, a *prescription* for agent 2 is a mapping $\Gamma_t^2 : \mathcal{L}_t^2 \to \mathcal{U}_t^2$ which takes values in a finite set $\mathcal{F}_t^2$.

At each t, the prescription for agent 2 is generated using a prescription law $\psi_t^2 : \mathcal{A}_t^2 \to \mathcal{F}_t^2$, which yields $\Gamma_t^2 = \psi_t^2(A_t^2)$. We call $\boldsymbol{\psi}^2 := (\psi_t^2 : t = 0, \dots, T)$ the prescription strategy for agent 2. Given a prescription Γ_t^2, the action of agent 2 is computed as $U_t^2 = \Gamma_t^2\big(L_t^2\big)$. Next, we use the person-by-person approach to set up a "new" centralized problem for agent 1. We proceed by arbitrarily fixing the prescription strategy $\boldsymbol{\psi}^2$ for agent 2. Since the prescription Γ_t^2 is generated using only the accessible information $A_t^2 \subseteq M_t^1$, agent 1 can derive the prescription using the fixed strategy as $\Gamma_t^2 = \psi_t^2(A_t^2)$. Then, we define a new state for agent 1 as $S_t^1 := \{X_t, L_t^2, A_t^2\}$ for all t, which takes values in a finite collection of sets $\mathcal{S}_t^1$. Given a prescription strategy $\boldsymbol{\psi}^2$, we can construct a state evolution function $\bar{f}_t^1(\cdot)$, such that $S_{t+1}^1 = \bar{f}_t^1(S_t^1, U_t^1, W_t, V_{t+1}^{1:2})$ and an observation rule $\bar{h}_t^1(\cdot)$ which yields $Z_{t+1}^1 = \bar{h}_t^1(S_t^1, U_t^1, W_t, V_{t+1}^{1:2})$ for all $t = 0, \dots, T-1$. The existence of these functions can be verified using the dynamics and information structure of the system to write the LHS in terms of the variables in the RHS. Similarly, we can construct a cost function $\bar{c}_t^1(\cdot)$ which yields the cost $\bar{c}_t^1(S_t^1, U_t^1) := c_t(X_t, U_t^1, \psi_t^2(A_t^2)(L_t^2))$ for all t. Then, for a given prescription strategy $\boldsymbol{\psi}^2$, the new centralized problem for agent 1 has state S_t^1, control action U_t^1, observation Z_t^1, and cost $\bar{c}_t^1(S_t^1, U_t^1)$ at time t. Furthermore, the performance criterion is $\mathcal{J}^1(\boldsymbol{g}^1) := \mathbb{E}^{\boldsymbol{g}^1}[\sum_{t=0}^T \bar{c}_t^1(S_t^1, U_t^1)]$.

Problem 2. The problem for agent 1 is $\inf_{\boldsymbol{g}^1} \mathcal{J}^1(\boldsymbol{g}^1)$, given a prescription strategy $\boldsymbol{\psi}^2$, the probability distributions of the primitive random variables $\{X_0, W_{0:t}, V_{0:t}^{1:2}\}$, and the dynamics $\{\bar{c}_t^1, \bar{f}_t^1, \bar{h}_t^1 : t = 0, \dots, T\}$.

Lemma 1. *For a given control strategy $\boldsymbol{g}^2$, consider a prescription strategy $\boldsymbol{\psi}^2$ such that*

$$\psi_t^2(A_t^2)(\cdot) := g_t^2(\cdot, A_t^2), \quad t = 0, \dots, T. \tag{4}$$

Then, $\mathcal{J}(\boldsymbol{g}^1, \boldsymbol{g}^2) = \mathcal{J}^1(\boldsymbol{g}^1)$ for the fixed prescription strategy $\boldsymbol{\psi}^2$. Moreover, for any given prescription strategy $\boldsymbol{\psi}^2$, consider a control strategy $\boldsymbol{g}^2$ constructed as

$$g_t^2(\cdot, A_t^2) := \psi_t^2(A_t^2)(\cdot), \quad t = 0, \dots, T. \tag{5}$$

Then, $\mathcal{J}^1(\boldsymbol{g}^1)$ after fixing $\boldsymbol{\psi}^2$ is equal to $\mathcal{J}(\boldsymbol{g}^1, \boldsymbol{g}^2)$.

Proof. For the first part, given a control strategy $\boldsymbol{g}$ and prescription strategy $\boldsymbol{\psi}^2$, note that $U_t^2 = g_t^2(L_t^2, A_t^2) = \psi_t^2(A_t^2)(L_t^2)$, i.e., the control law and prescription law result in the same control action U_t^2 for a given memory $M_t^2 = \{L_t^2, A_t^2\}$, for all $t = 0, \dots, T$. Thus, after fixing $\boldsymbol{\psi}^2$, we can write the expected cost at each t as $\mathbb{E}^{\boldsymbol{g}}[c_t(X_t, U_t^{1:2})] = \mathbb{E}^{\boldsymbol{g}^1}[c_t(X_t, U_t^1, \psi_t^2(A_t^2)(L_t^2))] = \mathbb{E}^{\boldsymbol{g}^1}[\bar{c}_t^1(S_t^1, U_t^1)]$, where the second equality holds using the construction of $\bar{c}_t^1(\cdot)$. The proof is complete by summing the cost over all time steps. For the second part, the proof follows from similar arguments as in the first part. □

Remark 3. We consider that a control strategy $\boldsymbol{g}^2$ and a prescription strategy $\boldsymbol{\psi}^2$ are always selected to satisfy (4) and (5) simultaneously. Thus, fixing $\boldsymbol{\psi}^2$ in Problem 2 also fixes $\boldsymbol{g}^2$, and vice versa. Next, consider a control strategy $(\boldsymbol{g}^{*1}, \boldsymbol{g}^{*2})$ which is an optimal solution to Problem 2. We construct a prescription strategy for agent 2 as $\psi_t^{*2}(A_t^2)(\cdot) := g_t^{*2}(\cdot, A_t^2)$, for all $t = 0, \dots, T$, and use the first part of Lemma 1 to conclude that $\boldsymbol{g}^{*1}$ must an optimal solution for Problem 2 after fixing $\boldsymbol{\psi}^{*2}$. Thus, every optimal solution to Problem 1 yields a corresponding solution to Problem 2.

Problem 2 is a centralized stochastic control problem for agent 1, with a perfectly observed component A_t^2 of the state S_t^1 and a partially observed component $\{X_t, L_t^2\}$, which must be estimated using the memory M_t^1. For such an estimation problem, it is known [4, page 79] that agent 1 can use the probability distribution

$$\Pi_t^1 := \mathbb{P}^{\boldsymbol{g}}\big(X_t, L_t^2 \mid M_t^1, \Gamma_{0:t-1}^2\big), \quad t = 0, \dots, T, \tag{6}$$

which takes values in the set of feasible distributions $\mathcal{P}_t^1 := \Delta(\mathcal{X}_t \times \mathcal{L}_t^2)$, where $\Gamma_{0:t-1}^2$ are known given $\boldsymbol{\psi}^2$ and M_t^1. Next, we show that the information state Π_t^1 evolves independent of the choice of strategies $\boldsymbol{g}^1$ and $\boldsymbol{\psi}^2$.

Lemma 2. *For all $t = 0, \dots, T-1$, there exists a function $\tilde{f}_t^1(\cdot)$ independent of control strategy $\boldsymbol{g}^1$ and prescription strategy $\boldsymbol{\psi}^2$, such that $\Pi_{t+1}^1 = \tilde{f}_t^1(\Pi_t^1, U_t^1, \Gamma_t^2, Z_{t+1}^1)$, and subsequently, for any Borel subset $P^1 \subseteq \mathcal{P}_{t+1}^1$, $\mathbb{P}(\Pi_{t+1}^1 \in P^1 | M_t^1, U_{0:t}^1, \Gamma_{0:t}^2) = \mathbb{P}(\Pi_{t+1}^1 \in P^1 | \Pi_t^1, U_t^1, \Gamma_t^2)$.*

Proof. The proof follows the same arguments as the ones of Lemma 4 in Section III-B. □

Lemma 3. *For any given prescription strategy $\boldsymbol{\psi}^2$ of agent 2, there exists a function $\tilde{c}_t^1(\cdot)$ for all $t = 0, \dots, T$, such that*

$$\mathbb{E}^{\boldsymbol{g}}[c_t(X_t, U_t^{1:2}) \mid M_t^1, U_t^1, \Gamma_t^2] = \tilde{c}_t^1(\Pi_t^1, A_t^2, U_t^1). \tag{7}$$

Proof. The proof follows the same arguments as the ones of Lemma 5 in Section III-B. □

The distribution Π_t^1 is called an *information state* of agent 1 at time t. As a consequence of Lemmas 2 and 3, the information state yields the following result for Problem 2.

Theorem 1. *For any given prescription strategy $\boldsymbol{\psi}^2$ of agent 2 in Problem 2, without loss of optimality, we can restrict attention to control strategies $\boldsymbol{g}^{*1}$ with the structural form*

$$U_t^1 = g_t^{*1}(A_t^2, \Pi_t^1), \quad t = 0, \dots, T. \tag{8}$$

Proof. This proof follows standard arguments for centralized stochastic control problems in [4, page 79], and thus, it is omitted. □

Theorem 1 establishes a structural form for an optimal control strategy $\boldsymbol{g}^{*1}$ in Problem 2, which holds for all $\boldsymbol{\psi}^2$, and subsequently, for all $\boldsymbol{g}^2$. From Remark 3, we note that any optimal control strategy $(\boldsymbol{g}^{*1}, \boldsymbol{g}^{*2})$ for Problem 1 must yield a corresponding prescription strategy $\boldsymbol{\psi}^{*2}$ such that after fixing $\boldsymbol{\psi}^{*2}$, the control strategy $\boldsymbol{g}^{*1}$ is the optimal solution to Problem 2. Thus, there exists an optimal control strategy $(\boldsymbol{g}^{*1}, \boldsymbol{g}^{*2})$ for Problem 1 where $\boldsymbol{g}^{*1}$ takes the structural form in (8).

Remark 4. Consider that $|\mathcal{X}_t \times \mathcal{L}_t^2| = m \in \mathbb{N}$. Then, the information state Π_t^1 takes values in the continuous space $\mathcal{P}_t^1 = \big\{\big(p_t(1), \dots, p_t(m)\big) \in [0,1]^m : \sum_{i=1}^m p_t(i) = 1\big\}$. However, for all $t = 0, \dots, T$, the information state can only take *countably* many realizations because all random variables take values in finite sets. For example, at $t = 0$, for each $x_0 \in \mathcal{X}_0$ and $l_0^2 \in \mathcal{L}_0^2$, the probability $\mathbb{P}^{\boldsymbol{g}}(x_0, l_0^2 \mid z_0^1)$ can take only finitely many values, i.e., one value for each $z_0^1 \in \mathcal{Z}_0^1$. Similarly, at any finite t, the memory M_t^1 can take finitely many realizations and thus, there are finitely many realizations for Π_t^1. As the horizon $T \to \infty$, the information state may take at most countably infinite realizations.

B. Analysis for Agent 2

In this subsection, we restrict agent 1 to control strategies $\boldsymbol{g}^1$ which satisfy (8), and derive a structural form for the optimal prescription strategy of agent 2. Given $\boldsymbol{g}^1$, agent 2 cannot generate the action U_t^1 at each t because they cannot access Π_t^1. Thus, we consider a two stage process to generate the action of agent 1: (1) agent 2 generates a prescription for agent 1 using only A_t^2, and (2) agent 1 computes U_t^1 using this prescription along with Π_t^1.

Definition 3. For all $t = 0, \dots, T$, a *prescription* for agent 1 is a function $\Gamma_t^1 : \mathcal{P}_t^1 \to \mathcal{U}_t^1$ which takes values in a finite set $\mathcal{F}_t^1$.

At each t, the prescription for agent 1 is generated using a prescription law $\psi_t^1 : \mathcal{A}_t^2 \to \mathcal{F}_t^1$, which yields $\Gamma_t^1 = \psi_t^1(A_t^2)$. We call $\boldsymbol{\psi}^1 := \big(\psi_t^1 : t = 0, \dots, T\big)$ the prescription strategy of agent 1 and $\boldsymbol{\psi} := (\boldsymbol{\psi}^1, \boldsymbol{\psi}^2)$ the prescription strategy of the system. For a given prescription Γ_t^1, agent 1 computes their action as $U_t^1 = \Gamma_t^1(\Pi_t^1)$. Next, we set up a new centralized problem from the perspective of agent 2 with a state $S_t^2 := \{X_t, L_t^2, \Pi_t^1\}$ for all t, which takes values in the finite collection of sets $\mathcal{S}_t^2$. Moreover, we can construct a state evolution function $\bar{f}_t^2(\cdot)$ such that $S_{t+1}^2 = \bar{f}_t^2(S_t^2, \Gamma_t^{1:2}, W_t, V_{t+1}^{1:2})$ and an observation rule $\bar{h}_t^2(\cdot)$ which yields $Z_{t+1}^2 = \bar{h}_t^2(S_t^2, \Gamma_t^{1:2}, W_t, V_{t+1}^{1:2})$ for all $t = 0, \dots, T-1$. Similarly, we can construct a cost function $\bar{c}_t^2(\cdot)$ such that $\bar{c}_t^2(S_t^2, \Gamma_t^{1:2}) := c_t(X_t, \Gamma_t^1(\Pi_t^1), \Gamma_t^2(L_t^2))$ for all t. Thus, the new centralized problem for agent 2 has the state S_t^2, observation Z_t^2 and action (Γ_t^1, Γ_t^2) at each t. The corresponding performance criterion is $\mathcal{J}^2(\boldsymbol{\psi}) = \mathbb{E}^{\boldsymbol{\psi}}[\sum_{t=0}^T \bar{c}_t^2(S_t^2, \Gamma_t^{1:2})]$.

Problem 3. The optimization problem for agent 2 is $\inf_{\boldsymbol{\psi}} \mathcal{J}^2(\boldsymbol{\psi})$, given the probability distributions of the primitive random variables $\{X_0, W_{0:t}, V_{0:t}^{1:2}\}$, and the dynamics $\{\bar{c}_t^2, \bar{f}_t^2, \bar{h}_t^2 : t = 0, \dots, T\}$.

Remark 5. Using the same sequence of arguments as Lemma 1, for each control strategy $\boldsymbol{g}$, we can construct an equivalent prescription strategy $\boldsymbol{\psi}$ such that $\mathcal{J}(\boldsymbol{g}) = \mathcal{J}(\boldsymbol{\psi})$ and vice versa. Thus, we always ensure that $\boldsymbol{\psi}$ is consistent with $\boldsymbol{g}$, which implies that for all t, $\Pi_t^1 = \mathbb{P}^{\boldsymbol{g}}(X_t \mid M_t^1, \Gamma_{0:t-1}^2) = \mathbb{P}^{\boldsymbol{\psi}}(X_t \mid M_t^1, \Gamma_{0:t-1}^2) = \mathbb{P}^{\boldsymbol{\psi}}(X_t \mid M_t^1, \Gamma_{0:t-1}^1, \Gamma_{0:t-1}^2)$, where we can add $\Gamma_{0:t-1}^1$ to the conditioning because they are functions of $A_t^2 \subseteq M_t^1$ and $\boldsymbol{\psi}^1$. Because of this property, we can equivalently write the dependence of a probability distribution on either $\boldsymbol{g}$ or $\boldsymbol{\psi}$.

Problem 3 is a partially observed centralized stochastic control problem and thus, agent 2 must estimate the state S_t^2 at each time t. For this purpose, agent 2 can use the distribution

$$\Pi_t^2 := \mathbb{P}^{\boldsymbol{\psi}}(X_t, L_t^2, \Pi_t^1 \mid A_t^2, \Gamma_{0:t-1}^{1:2}), \quad t = 0, \dots, T, \tag{9}$$

which takes values in the set of feasible distributions $\mathcal{P}_t^2 := \Delta(\mathcal{X}_t \times \mathcal{L}_t^2 \times \mathcal{P}_t^1)$. Recall that at each t, the information state of agent 1, Π_t^1, can take at most countably infinitely many realizations in the space $\mathcal{P}_t^1$. Thus, the information state Π_t^2 can be represented using a tuple of probability mass functions $\big(p_t(x_t, \ell_t^2, \cdot \mid a_t^2, \gamma_{0:t-1}^{1:2}) : x_t \in \mathcal{X}_t, \ell_t^2 \in \mathcal{L}_t^2\big)$, where $p_t(x_t, \ell_t^2, \cdot \mid a_t^2, \gamma_{0:t-1}^{1:2}) : \mathcal{P}_t^1 \to [0,1]$ for each $x_t \in \mathcal{X}_t$ and $\ell_t^2 \in \mathcal{L}_t^2$. Next, we show that the evolution of Π_t^2 is Markovian and independent of the prescription strategy $\boldsymbol{\psi}$.

Lemma 4. *For all $t = 0, \dots, T-1$, there exists a function $\tilde{f}_t^2(\cdot)$ independent of the prescription strategy $\boldsymbol{\psi}$, such that $\Pi_{t+1}^2 = \tilde{f}_t^2(\Pi_t^2, \Gamma_t^1, \Gamma_t^2, Z_{t+1}^2)$, and subsequently, for any Borel subset $P^2 \subseteq \mathcal{P}_{t+1}^2$, $\mathbb{P}(\Pi_{t+1}^2 \in P^2 \mid A_t^2, \Gamma_{0:t}^{1:2}) = \mathbb{P}(\Pi_{t+1}^2 \in P^2 \mid \Pi_t^2, \Gamma_t^{1:2})$.*

Proof. Let x_t, γ_t^1, γ_t^2, a_t^2, and π_t^1 be realizations of X_t, Γ_t^1, Γ_t^2, A_t^2, and the distribution Π_t^1, respectively, for all t. Then, using Bayes' rule

$$\begin{aligned}\mathbb{P}^{\boldsymbol{\psi}}&(x_{t+1}, \ell_{t+1}^2, \pi_{t+1}^1 \mid a_{t+1}^2, \gamma_{0:t}^{1:2}) \\ &= \frac{\mathbb{P}^{\boldsymbol{\psi}}\big(x_{t+1}, \ell_{t+1}^2, \pi_{t+1}^1, z_{t+1}^2 \mid a_t^2, \gamma_{0:t}^{1:2}\big)}{\mathbb{P}^{\boldsymbol{\psi}}\big(z_{t+1}^2 \mid a_t^2, \gamma_{0:t}^{1:2}\big)},\end{aligned} \tag{10}$$

where $a_{t+1}^2 = a_t^2 \cup z_{t+1}^2$. Using the dynamics $\{\bar{f}_t^2, \bar{h}_t^2, \bar{c}_t^2\}$, we write that $(x_{t+1}, \ell_{t+1}^2) = \eta_t^2(s_t^2, \gamma_t^{1:2}, w_t, v_{t+1}^{1:2})$, $\pi_{t+1}^1 = \xi_t^2(s_t^2, \gamma_t^{1:2}, w_t, v_{t+1}^{1:2})$, $z_{t+1}^2 = \bar{h}_t^2(s_t^2, \gamma_t^{1:2}, w_t, v_{t+1}^{1:2})$, for some appropriate functions $\eta_t^2(\cdot)$ and $\xi_t^2(\cdot)$, where $s_t^2 = \{x_t, \ell_t^2, \pi_t^1\}$. Substituting these relationships into the numerator in the RHS of (10) yields that

$$\begin{aligned}&\mathbb{P}^{\boldsymbol{\psi}}\big(x_{t+1}, \ell_{t+1}^2, \pi_{t+1}^1, z_{t+1}^2 \mid a_t^2, \gamma_{0:t}^{1:2}\big) \\ &= \sum_{s_t^2, w_t, v_{t+1}^{1:2}} \mathbb{I}[\eta_t^2(s_t^2, \gamma_t^{1:2}, w_t) = (x_{t+1}, \ell_{t+1}^2)] \cdot \mathbb{P}(w_t, v_{t+1}^{1:2}) \\ &\quad \cdot \mathbb{I}[\xi_t^2(s_t^2, \gamma_t^{1:2}, w_t, v_{t+1}^{1:2}) = \pi_{t+1}^1] \cdot \mathbb{P}^{\boldsymbol{\psi}}\big(s_t^2 \mid a_t^2, \gamma_{0:t-1}^{1:2}\big) \\ &\qquad \cdot \mathbb{I}[\bar{h}_{t+1}^2(s_t^2, \gamma_t^{1:2}, w_t, v_{t+1}^{1:2}) = z_{t+1}^2],\end{aligned} \tag{11}$$

where $\mathbb{I}(\cdot)$ is the indicator function, and where we can drop the prescriptions $\gamma_t^{1:2}$ from the conditioning in the last term because they are completely determined given $\boldsymbol{\psi}$ and a_t^2. Note that in (11), $\mathbb{P}^{\boldsymbol{\psi}}\big(s_t^2 \mid a_t^2, \gamma_{0:t-1}^{1:2}\big) = \pi_t^2(s_t^2)$. Next, we expand the denominator in (10) as

$$\mathbb{P}^{\boldsymbol{\psi}}\big(z_{t+1}^2 \mid a_t^2, \gamma_{0:t}^{1:2}\big) = \sum_{s_t^2, w_t, v_{t+1}^{1:2}} \mathbb{P}(w_t, v_{t+1}^{1:2}) \\ \cdot \mathbb{I}[\bar{h}_t^2(s_t^2, \gamma_t^{1:2}, w_t, v_{t+1}^{1:2}) = z_{t+1}^2] \cdot \pi_t^2(s_t^2). \quad (12)$$

Then, the first result holds by constructing an appropriate function $\tilde{f}_t^2(\cdot)$ using (10) - (12). To prove the second result, for any Borel subset $P^2 \subseteq \mathcal{P}_{t+1}^2$, we write that

$$\mathbb{P}(\Pi_{t+1}^2 \in P^2 \mid a_t^2, \gamma_{0:t}^{1:2}, \pi_{0:t}^2) = \sum_{z_{t+1}^2} \mathbb{I}[\tilde{f}_t^2(\pi_t^2, \gamma_t^{1:2}, \\ z_{t+1}^2) \in P^2] \cdot \mathbb{P}(z_{t+1}^2 \mid a_t^2, \gamma_{0:t}^{1:2}, \pi_{0:t}^2). \quad (13)$$

The second term in (13) can be expanded as $\mathbb{P}(z_{t+1}^2 \mid a_t^2, \gamma_{0:t}^{1:2}, \pi_{0:t}^2) = \sum_{s_t^2, w_t, v_{t+1}^{1:2}} \mathbb{I}[\bar{h}_t^2(s_t^2, \gamma_t^{1:2}, w_t, v_{t+1}^{1:2}) = z_{t+1}^2] \cdot \mathbb{P}(v_{t+1}^{1:2}, w_t) \cdot \pi_t^2(s_t^2)$. The proof is complete by substituting this equation into (13). □

Lemma 5. *There exists a function $\tilde{c}_t^2(\cdot)$ for all t, such that*

$$\mathbb{E}^{\boldsymbol{g}}[c_t(X_t, U_t^{1:2}) \mid A_t^2, \Gamma_t^{1:2}] = \tilde{c}_t^2(\Pi_t^2, \Gamma_t^{1:2}). \quad (14)$$

Proof. Let a_t^3, $\gamma_t^{1:2}$, and π_t^2 be realizations of the random variables A_t^3, $\Gamma_t^{1:2}$, and the conditional distribution Π_t^2, respectively, for all $t = 0, \dots, T$. To prove the result, we expand the expectation as $\mathbb{E}^{\boldsymbol{g}}[c_t(X_t, U_t^{1:2}) \mid a_t^2, \gamma_t^{1:2}] = \mathbb{E}^{\boldsymbol{\psi}}[\bar{c}_t^2(S_t^2, \Gamma_t^{1:2}) \mid a_t^2, \gamma_t^{1:2}] = \sum_{s_t^2} \bar{c}_t^2(s_t^2, \gamma_t^{1:2}) \cdot \mathbb{P}^{\boldsymbol{\psi}}(S_t^2 = s_t^2 \mid a_t^2, \gamma_t^{1:2}) = \sum_{s_t^2} \bar{c}_t^2(s_t^2, \gamma_t^{1:2}) \cdot \pi_t^2(s_t^2) =: \tilde{c}_t^2(\pi_t^2, \gamma_t^{1:2})$, where we can drop the prescripitions $\gamma_t^{1:2}$ from the conditioning because they known given $\boldsymbol{\psi}$ and a_t^2. □

We call Π_t^2 the information state of agent 2 at time t. As a consequence of Lemmas 2 and 3, the information state yields the following result for Problem 3.

Theorem 2. *In Problem 3, without loss of optimality, we can restrict our attention to prescription strategies $\boldsymbol{\psi}^*$ with the structural form*

$$\Gamma_t^k = \psi_t^{*k}(\Pi_t^2), \quad k = 1, 2, \quad t = 0, \dots, T. \quad (15)$$

Proof. This proof follows similar arguments for centralized stochastic control problems in [4, page 79], and thus, it is omitted. □

Consider a prescription strategy $\boldsymbol{\psi}^* = (\boldsymbol{\psi}^{*1}, \boldsymbol{\psi}^{*2})$ which is an optimal solution to Problem 3, and a control strategy $\boldsymbol{g}^*) = \boldsymbol{g}^{*1}, \boldsymbol{g}^{*2})$ given by $g_t^{*1}(\Pi_t^{1:2}) := \psi_t^{*1}(\Pi_t^2)(\Pi_t^1)$ and $g_t^{*2}(L_t^2, \Pi_t^2) := \psi_t^{*2}(\Pi_t^2)(L_t^2)$ for each $k = 1, 2$ and $t = 0, \dots, T$. Using the same arguments as in Lemma 1, we conclude that $\mathcal{J}(\boldsymbol{g}^*) = \mathcal{J}^2(\boldsymbol{\psi}^*)$ and subsequently, that $\boldsymbol{g}^*$ is the optimal solution to Problem 1. Thus, without loss of optimality, we can restrict attention to control strategies $\boldsymbol{g}^*$ with the structural form $U_t^1 = g_t^{*1}(\Pi_t^1, \Pi_t^2)$ and $U_t^2 = g_t^{*2}(L_t^2, \Pi_t^2)$ for all $t = 0, \dots, T$.

Remark 6. Consider a system where, the feasible sets of system variables are time invariant, i.e., $\mathcal{X}_t = \mathcal{X}$, $\mathcal{W}_t = \mathcal{W}$, $\mathcal{V}_t^k = \mathcal{V}^k$, $\mathcal{Y}_t^k = \mathcal{Y}^k$ for each $k = 1, 2$ and $t = 0, \dots, T$, and the information structure satisfies $\mathcal{L}_t^2 = \mathcal{L}^2$, $\mathcal{Z}_t^1 = \mathcal{Z}^1$, $\mathcal{Z}_t^2 = \mathcal{Z}^2$ for all t. Note that the set $\mathcal{M}_t^1$ still grows in size with time. However, the spaces $\mathcal{P}^1 = \Delta(\mathcal{X} \times \mathcal{L}^2)$ and $\mathcal{P}^2 = \Delta(\mathcal{X} \times \mathcal{L}^2 \times \mathcal{P}^1)$ are time invariant and subequently, our optimal control strategies have time-invariant domains for both agents. This is a useful property to derive and implement optimal control strategies for long time horizons.

C. Dynamic Programming Decomposition

In this subsection, we construct the value functions and corresponding control laws to form a DP decomposition which can derive the optimal prescription strategies. Let γ_t^k and π_t^k be the realizations of the prescription Γ_t^k and information state Π_t^k, respectively, for each $k = 1, 2$ and $t = 0, \dots, T$. Then, we recursively define the value functions

$$J_t(\pi_t^2) := \inf_{\gamma_t^{1:2} \in \mathcal{F}_t^1 \times \mathcal{F}_t^2} \tilde{c}_t^2\big(\pi_t^2, \gamma_t^{1:2}\big) \\ + \mathbb{E}^{\boldsymbol{\psi}}\Big[J_{t+1}\big(\tilde{f}_t^2(\pi_t^2, \gamma_t^{1:2}, Z_{t+1}^2)\big) \mid \pi_t^2, \gamma_t^{1:2}\Big], \quad (16)$$

for all $t = 0, \dots, T$ and define $J_{T+1}(\pi_{T+1}^2) := 0$ identically. For each agent $k = 1, 2$, the prescription law at time t is $\gamma_t^{*k} = \psi_t^{*k}(\pi_t^2)$, i.e., the $\arg\inf$ in the RHS of (16). The prescription strategy $\boldsymbol{\psi}^*$ derived using this DP decomposition can be shown to be the optimal solution to Problem 2 using standard arguments [6], [18]. Recall that given an optimal strategy $\boldsymbol{\psi}^*$ derived using this DP decomposition, we can also derive the optimal control strategy $\boldsymbol{g}^*$ for Problem 1.

Remark 7. At each $t = 0, \dots, T$, our DP decomposition requires solving an optimization problem for each realization π_t^2 of the information state Π_t^2, which is a tuple of probability mass functions. Optimizing over probability mass functions is a computationally challenging problem. Next, we present two different approaches to alleviate the computational implications. In Section IV, we show how we can simplify our results when the system dynamics and information structure have additional favorable properties. In Section V, we present an approximation for the information states which can reduce the number of computations required to derive an approximately optimal strategy.

IV. Simplification for Decoupled Dynamics

In this subsection, we show how our results can be simplified when both agents have decoupled state and observation dynamics. We denote the state of each agent $k = 1, 2$ at time t by $X_t^k \in \mathcal{X}_t^k$. Starting at X_0^k, each state evolves as

$$X_{t+1}^k = f_t^k(X_t^k, U_t^k, W_t^k), \quad t = 0, \dots, T-1, \quad (17)$$

for $k = 1, 2$, where $W_t^k \in \mathcal{W}_t^k$ is a disturbance acting only on X_t^k. The observation of agent k at time t is $Y_t^k = h_t^k(X_t^k, V_t^k)$. We assume that all primitive random variables $\{X_0^k, W_t^k, V_t^k : k = 1, 2, \ t = 0, \dots, T\}$ are independent of each other and that the cost to the system at each $t = 0, \dots, T$ is $c_t(X_t^{1:2}, U_t^{1:2}) \in \mathbb{R}_{\geq 0}$. Without loss of optimality, we restrict attention to control strategies where $\boldsymbol{g}^1$ takes the form $U_t^1 = g_t^1(\Pi_t^1, \Pi_t^2)$ and where $\boldsymbol{g}^2$ takes the form $U_t^2 = g_t^2(L_t^2, \Pi_t^2)$, for all $t = 0, \dots, T$. Here, recall that $\Pi_t^1 = \mathbb{P}^{\boldsymbol{g}}\big(X_t^{1:2}, L_t^2 | M_t^1, \Gamma_{0:t-1}^2\big)$ and $\Pi_t^2 = \mathbb{P}^{\boldsymbol{g}}\big(X_t^{1:2}, L_t^2, \Pi_t^1 | A_t^2, \Gamma_{0:t-1}^{1:2}\big)$. Next, we show that the information state Π_t^1 can be simplified using the decoupled dynamics.

Lemma 6. *For each $k = 1, 2$ and $t = 0, \dots, T$, let x_t^k, m_t^k, l_t^2, and a_t^2 be realizations of the random variables X_t^k, M_t^k, L_t^2, and A_t^2, respectively. Then,*

$$\mathbb{P}^{\boldsymbol{g}}(x_t^{1:2}, l_t^2 \mid m_t^1) = \mathbb{P}^{\boldsymbol{g}}(x_t^1 \mid m_t^1) \cdot \mathbb{P}^{\boldsymbol{g}}(x_t^2, l_t^2 \mid a_t^2). \quad (18)$$

Proof. Given the realizations x_t^k, y_t^k, u_t^k, γ_t^k and l_t^2 of X_t^k, Y_t^k, U_t^k, Γ_t^k, and L_t^2, respectively, for each $k = 1, 2$ and $t = 0, \dots, T$, we prove (18) by mathematical induction. At $t = 0$, depending on the information sharing pattern of the system, there are two possible realizations of the memory of agent 1, either $m_0^1 = \{y_0^1\}$ or $m_0^1 = \{y_0^1, y_0^2\}$. For the first realization of the memory of agent 1, the private information of agent 2 is $l_0^2 = \{y_0^2\}$, and thus, we can expand the LHS of (18) as $\mathbb{P}^{\boldsymbol{g}}(x_0^{1:2}, y_0^2 | m_0^1) = \mathbb{P}^{\boldsymbol{g}}(x_0^{1:2}, y_0^2 | y_0^1) = \mathbb{P}^{\boldsymbol{g}}(x_0^1 | y_0^1) \cdot \mathbb{P}^{\boldsymbol{g}}(x_0^2, y_0^2)$, where recall that the observation y_0^k depends only on x_0^k for each k, and the primitive random variables are independent of each other. For the second realization of the memory of agent 1, note that $l_t^2 = \emptyset$ because $l_t^2 \cap m_t^1 = \emptyset$, and thus, we can expand the LHS as $\mathbb{P}^{\boldsymbol{g}}(x_0^{1:2} | m_0^1) = \mathbb{P}^{\boldsymbol{g}}(x_0^1 | y_0^1) \cdot \mathbb{P}^{\boldsymbol{g}}(x_0^2 | y_0^2)$. For both cases at $t = 0$, we have shown the LHS is equal to the RHS in (18). This forms the basis of our induction. Next, we consider the induction hypothesis that (18) holds at each $0, \dots, t$, and expand the LHS at $t + 1$ as

$$\begin{aligned}\mathbb{P}^{\boldsymbol{g}}(x_{t+1}^{1:2}, l_{t+1}^2 \mid m_{t+1}^1) &= \frac{\mathbb{P}^{\boldsymbol{g}}(x_{t+1}^{1:2}, l_{t+1}^2, z_{t+1}^1 \mid m_t^1)}{\mathbb{P}^{\boldsymbol{g}}(z_{t+1}^1 \mid m_t^1)} \\ &= \frac{\mathbb{P}^{\boldsymbol{g}}(x_{t+1}^{1:2}, l_{t+1}^2, z_{t+1}^1 \mid m_t^1)}{\sum_{x_{t+1}^{1:2}, l_{t+1}^2} \mathbb{P}^{\boldsymbol{g}}(x_{t+1}^{1:2}, l_{t+1}^2, z_{t+1}^1 \mid m_t^1)}. \quad (19)\end{aligned}$$

Note that in the partially accessible information structure, $l_{t+1}^2 \cup z_{t+1}^1 = l_t^2 \cup \{y_{t+1}^{1:2}, u_t^{1:2}\}$. Thus, we can write that $\mathbb{P}^{\boldsymbol{g}}(x_{t+1}^{1:2}, l_{t+1}^2, z_{t+1}^1 \mid m_t^1) = \mathbb{P}^{\boldsymbol{g}}(x_{t+1}^{1:2}, y_{t+1}^{1:2}, u_t^{1:2}, l_t^2 \mid m_t^1) = \mathbb{P}^{\boldsymbol{g}}(y_{t+1}^1 | x_{t+1}^1) \cdot \mathbb{P}^{\boldsymbol{g}}(y_{t+1}^2 | x_{t+1}^2) \cdot \mathbb{I}[g_t^1(m_t^1) = u_t^1] \cdot \mathbb{I}[\gamma_t^2(l_t^2) = u_t^2] \cdot \mathbb{P}^{\boldsymbol{g}}(x_{t+1}^{1:2}, l_t^2 | m_t^1)$, where $\mathbb{I}(\cdot)$ is the indicator function, and where γ_t^2 and u_t^1 are completely determined given m_t^1 and $\boldsymbol{g}$. Furthermore, we expand the last term as $\mathbb{P}^{\boldsymbol{g}}(x_{t+1}^{1:2}, l_t^2 | m_t^1, u_t^1, \gamma_t^2) = \sum_{x_t^{1:2}, w_t^{1:2}} \mathbb{I}[f_t^1(x_t^1, u_t^1, w_t^1) = x_{t+1}^1] \cdot \mathbb{I}[f_t^2(x_t^2, \gamma_t^2(l_t^2), w_t^2) = x_{t+1}^2] \cdot \mathbb{P}(w_t^1, w_t^2) \cdot \mathbb{P}^{\boldsymbol{g}}(x_t^{1:2}, l_t^2 | m_t^1)$, where we can use the induction hypothesis to obtain $\mathbb{P}^{\boldsymbol{g}}(x_t^{1:2}, l_t^2 | m_t^1) = \mathbb{P}^{\boldsymbol{g}}(x_t^1 | m_t^1) \cdot \mathbb{P}^{\boldsymbol{g}}(x_t^{1:2}, l_t^2 | a_t^2)$. Substituting these results into (19), and rearranging the terms yields $\mathbb{P}^{\boldsymbol{g}}(x_{t+1}^{1:2}, l_{t+1}^2 | m_{t+1}^1) = \frac{\mathbb{P}^{\boldsymbol{g}}(x_{t+1}^1, y_{t+1}^1, u_t^1 | m_t^1)}{\mathbb{P}^{\boldsymbol{g}}(y_{t+1}^1, u_t^1 | m_t^1)} \cdot \mathbb{P}^{\boldsymbol{g}}(x_{t+1}^2, l_{t+1}^2 | a_t^2, z_{t+1}^2) = \mathbb{P}^{\boldsymbol{g}}(x_{t+1}^1 | m_t^1, y_{t+1}^1, u_t^1) \cdot \mathbb{P}^{\boldsymbol{g}}(x_{t+1}^2, y_{t+1}^2 | a_{t+1}^2)$. To complete the proof by mathematical induction, we need to show that the first term in the RHS of the previous equation is equal to the first term in the RHS of (18). We achieve this by expanding $\mathbb{P}^{\boldsymbol{g}}(x_{t+1}^1 | m_t^1, y_{t+1}^{1:2}, u_t^{1:2}, l_t^2) = \frac{\mathbb{P}^{\boldsymbol{g}}(x_{t+1}^1, y_{t+1}^1 | m_t^1, l_t^2, y_{t+1}^2)}{\sum_{x_{t+1}^1} \mathbb{P}^{\boldsymbol{g}}(x_{t+1}^1, y_{t+1}^1 | m_t^1, l_t^2, y_{t+1}^2)}$ $= \frac{\sum_{x_t^1} \mathbb{P}^{\boldsymbol{g}}(y_{t+1}^1 | x_{t+1}^1) \cdot \mathbb{P}^{\boldsymbol{g}}(x_{t+1}^1 | x_t^1, u_t^1) \cdot \mathbb{P}^{\boldsymbol{g}}(x_t^1 | m_t^1)}{\sum_{x_{t:t+1}^1} \mathbb{P}^{\boldsymbol{g}}(y_{t+1}^1 | x_{t+1}^1) \cdot \mathbb{P}^{\boldsymbol{g}}(x_{t+1}^1 | x_t^1, u_t^1) \cdot \mathbb{P}^{\boldsymbol{g}}(x_t^1 | m_t^1)}$, where, in the last equality, we use Bayes' rule and the induction hypothesis. Recall that $z_{t+1}^1 \subseteq l_t^2 \cup \{y_{t+1}^{1:2}, u_t^{1:2}\}$. This implies that $\mathbb{P}^{\boldsymbol{g}}(x_{t+1}^1 | m_t^1, y_{t+1}^{1:2}, u_t^{1:2}, l_t^2) = \mathbb{P}^{\boldsymbol{g}}(x_{t+1}^1 | m_t^1, z_{t+1}^1) = \mathbb{P}^{\boldsymbol{g}}(x_{t+1}^1 | m_t^1, y_{t+1}^1, u_t^1)$, which complete the proof. □

Motivated by Lemma 6, we define the distributions $\Theta_t^1 := \mathbb{P}^{\boldsymbol{g}}(X_t^1 | M_t^1)$ and $\Theta_t^2 := \mathbb{P}^{\boldsymbol{g}}(X_t^2, L_t^2 | A_t^2)$ and note that the information state Π_t^1 at each $t = 0, \dots, T$ can be written as a function of (Θ_t^1, Θ_t^2). Thus, at time t, agent 1 can track the distributions (Θ_t^1, Θ_t^2) instead of Π_t^1 to compute their optimal control action U_t^1. Next, we show that the evolution of Θ_t^k, for each $k = 1, 2$, is Markovian, strategy independent and decoupled from the dynamics of the other agent.

Lemma 7. *At each time t, there exists a function $\tilde{e}_t^k(\cdot)$, independent of the strategy $\boldsymbol{g}$, for all $k = 1, 2$ such that*

$$\Theta_{t+1}^k = \tilde{e}_t^k(\Theta_t^k, U_t^k, Y_{t+1}^k). \quad (20)$$

Proof. The proof follows the same arguments as the ones in Lemma 4 and thus, due to space limitations, it is omitted. □

Note that the distribution Θ_t^2 is also available to agent 2 at each $t = 0, \dots, T$, because it depends only on the accessible information A_t^2. Subsequently, using the same sequence of arguments as the ones in Theorem 2, we conclude that, without loss of optimality, agent 2 can restrict attention to prescription strategies with the structural form $\Gamma_t^k = \psi_t^k\big(\mathbb{P}^{\boldsymbol{g}}(X_t^{1:2}, L_t^2, \Theta_t^1 | A_t^2), \Theta_t^2\big)$, for each $k = 1, 2$ and $t = 0, \dots, T$. Next, we show that the term $\mathbb{P}^{\boldsymbol{g}}(X_t^{1:2}, L_t^2, \Theta_t^1 | A_t^2)$ in the argument of the prescription law for each k can also be simplified using the decoupled dynamics of the system.

Lemma 8. *For each $k = 1, 2$ and $t = 0, \dots, T$, let x_t^k, l_t^2, a_t^2, and θ_t^k be realizations of the random variables X_t^k, L_t^2, A_t^2, and the probability distribution Θ_t^k, respectively. Then,*

$$\mathbb{P}^{\boldsymbol{g}}(x_t^{1:2}, l_t^2, \theta_t^1 | a_t^2) = \mathbb{P}^{\boldsymbol{g}}(x_t^1, \theta_t^1 | a_t^2) \cdot \mathbb{P}^{\boldsymbol{g}}(x_t^2, l_t^2 | a_t^2). \quad (21)$$

Proof. The proof follows by mathematical induction using the same arguments as the ones in Lemma 6, and thus, due to space limitations, it is omitted. □

Starting with the structural form of optimal prescription strategies in Theorem 2, we can use Lemmas 6 and 8, to conclude that in systems with decoupled dynamics, without loss of optimality, we can restrict attention to control strategies $\boldsymbol{g}^*$ with the structural form

$$\begin{aligned}U_t^1 &= g_t^{*1}\big[\Theta_t^1, \Theta_t^2, \mathbb{P}^{\boldsymbol{g}}(X_t^1, \Theta_t^1 \mid A_t^2)\big], & (22) \\ U_t^2 &= g_t^{*2}\big[L_t^2, \Theta_t^2, \mathbb{P}^{\boldsymbol{g}}(X_t^1, \Theta_t^1 \mid A_t^2)\big],\ t = 0, \dots, T. & (23)\end{aligned}$$

Remark 8. The control strategy $\boldsymbol{g}^1$ yielded a control law for each $t = 0, \dots, T$ for agent 1 with the form $U_t^1 = g_t^1(\Pi_t^1, \Pi_t^2)$, which has the domain $\Delta(\mathcal{X}_t^1 \times \mathcal{X}_t^2 \times \mathcal{L}_t^2) \times \Delta(\mathcal{X}_t^1 \times \mathcal{X}_t^2 \times \mathcal{L}_t^2 \times \Delta(\mathcal{X}_t^1 \times \mathcal{X}_t^2 \times \mathcal{L}_t^2))$. In contrast, the domain of the control law g_t^{*1} in (22) is $\Delta(\mathcal{X}_t^1) \times \Delta(\mathcal{X}_t^2 \times \mathcal{L}_t^2) \times \Delta(\mathcal{X}_t^1 \times \Delta(\mathcal{X}_t^1))$, which is a space with a smaller dimension than the one before. Similarly, the control laws of agent 2 have a domain with a smaller dimension in (23) than the control laws derived using Theorem 2. Thus, we have obtained a simpler form for an optimal control strategy in systems with decoupled dynamics.

We can further simplify the structural form of the optimal control strategies when agent 1 can perfectly observe the state X_t^1, i.e, $Y_t^1 = X_t$ and subsequently, $X_t^1 \subseteq M_t^1$ at each $t = 0, \ldots, T$. Then, for a given realization m_t^1 of the memory M_t^1, the probability distribution Θ_t^1 at each t is simply given by $\Theta_t^1 = \mathbb{I}[X_t^1 = x_t^1]$ for the realization $x_t^1 \in m_t^1$ of X_t^1, where $\mathbb{I}$ is the indicator function. Using this result in (22) and (23), we conclude that, without loss of optimality, we can restrict attention to control strategies $\boldsymbol{g}^*$ with the form

$$U_t^1 = g_t^1\big[X_t^1, \Theta_t^2, \mathbb{P}^{\boldsymbol{g}}(X_t^1 \mid A_t^2)\big], \tag{24}$$

$$U_t^2 = g_t^2\big[L_t^2, \Theta_t^2, \mathbb{P}^{\boldsymbol{g}}(X_t^1 \mid A_t^2)\big], \quad t = 0, \ldots, T. \tag{25}$$

Remark 9. When agent 1 can perfectly observe their own state, at each t, the domains of the optimal control laws g_t^{*1} and g_t^{*2} are $\mathcal{X}_t^1 \times \Delta(\mathcal{X}_t^2 \times \mathcal{L}_t^2) \times \Delta(\mathcal{X}_t^1)$ and $\mathcal{L}_t^2 \times \Delta(\mathcal{X}_t^2 \times \mathcal{L}_t^2) \times \Delta(\mathcal{X}_t^1)$, respectively. These domains are small enough that the optimal control laws at each t are functions of distributions over finite sets instead of probability mass functions. Thus, the resulting DP can be solved using standard techniques for centralized problems.

V. Implementation

In this subsection, we present an approach to approximate the information state Π_t^1 for all $t = 0, \ldots, T$ which ensures that the approximation can only take finitely many values. To simplify the notation, we restrict our attention to systems where $|\mathcal{X}_t \times \mathcal{L}_t^2| = m$, $m \in \mathbb{N}$ for all $t = 0, \ldots, T$. Furthermore, we consider that the maximum cost at each t is bounded above by $||c||_\infty < \infty$. Recall that the space of feasible values for Π_t^1 is the simplex $\mathcal{P}^1 = \big\{\big(p(1), \ldots, p(m)\big) \in [0,1]^m : \sum_{i=1}^m p(i) = 1\big\}$. We use the procedure in [19] to generate a set of equally distributed points in $\mathcal{P}^1$. Specifically, we select a number $n \in \mathbb{N}$ and define a set $\mathcal{Q}_n := \big\{\big(q(1), \ldots, q(m)\big) \in \mathcal{P}^1 : n \cdot q(i) \in \mathbb{N}_{\geq 0}, i = 1, \ldots, m\big\}$. The set $\mathcal{Q}_n$ forms a lattice containing $|\mathcal{Q}_n| = \binom{m+n-1}{m-1}$ points in the simplex $\mathcal{P}^1$. For example, let $\mathcal{X}_t = \{0, 1\}$ and $\mathcal{L}_t^2 = \emptyset$, which implies that $m = 2$. Then, by selecting $n = 2$ we construct the set $\mathcal{Q}_2 = \big\{(0,1), (1/2, 1/2), (1, 0)\big\}$. Similarly, if $m = 3$ and we select $n = 2$, we construct the set $\mathcal{Q}_2 = \big\{(1,0,0), (1/2,1/2,0), (0,1,0), (0,1/2,1/2), (0,0,1), (1/2,0,1/2)\big\}$. Next, we define the total variation distance between any point in $\mathcal{P}^1$ and $\mathcal{Q}_n$, and then, we use this metric to define an approximate information state.

Definition 4. The total variation distance between any $\pi_t^1 = (p(1), \ldots, p(m)) \in \mathcal{P}^1$ and any $q_t = (q(1), \ldots, q(m)) \in \mathcal{Q}_n$ is $|\pi_t^1 - q|_{TV} = \sum_{i=1}^m |p(i) - q(i)|$.

Definition 5. The approximate information state for agent 1 at each $t = 0, \ldots, T$ is a random variable $\hat{\Pi}_t^1$ which takes values in the finite set $\mathcal{Q}_n$, and which is given by

$$\hat{\Pi}_t^1 = \sigma(\Pi_t^1) := \arg\min_{q \in \mathcal{Q}_n} |\Pi_t^1 - q|_{TV}. \tag{26}$$

Given any distribution $\pi_t^1 \in \mathcal{P}^1$, the corresponding realization of the approximate information state $\hat{\pi}_t^1 = \sigma(\pi_t^1)$ can be efficiently computed using the algorithm in [19]. Next, we present an upper bound in the total variation distance between any information state and its approximation.

Lemma 9. *For all $t = 0, \ldots, T$, for any realization π_t^1 of the information state Π_t^1, it holds that $|\pi_t^1 - \sigma(\pi_t^1)|_{TV} \leq \frac{2a \cdot (1+a)}{m \cdot n}$, where $a = \lfloor m/2 \rfloor \in \mathbb{N}$ and $\lfloor \cdot \rfloor$ is the floor function.*

Proof. The proof follows from [19, Proposition 2]. □

Given any upper bound $\epsilon \in \mathbb{R}_{>0}$, we can use Lemma 9 to construct a set $\mathcal{Q}_n$ which satisfies $|\pi_t^1 - \sigma(\pi_t^1)| \leq \epsilon$ for all $\pi_t^1 \in \mathcal{P}^1$, by selecting $n \geq \frac{2a(1+a)}{m \cdot \epsilon}$. Furthermore, the resulting approximate information state $\hat{\Pi}_t^1$ can be updated in a Markovian and strategy independent manner as $\hat{\Pi}_{t+1}^1 = \sigma[\tilde{f}_t^1(\hat{\Pi}_t^1, U_t^1, \Gamma_t^2, Z_{t+1}^1)]$, for all $t = 0, \ldots, T-1$.

Our aim is to solve the centralized Problem 2 for agent 1 using the approximate information state $\hat{\Pi}_t^1$, which takes only finitely many values for all $t = 0, \ldots, T$, instead of the information state Π_t^1, which can take countably infinitely many values. For a fixed prescription strategy $\boldsymbol{\psi}^2$, recall from Lemma 3 that the expected cost at time t can be written as $\tilde{c}_t^1(\Pi_t^1, A_t^2, U_t^1)$. Then, in Problem 2, we can optimize the performance criterion $\mathcal{J}^1(\boldsymbol{g}^1)$ using a centralized DP as follows. Let u_t^1, a_t^2, and π_t^1 be the realizations of U_t^1, A_t^2 and Π_t^1, respectively. Then, we define the value functions

$$\begin{aligned} J_t^1(\pi_t^1, a_t^2) &:= \inf_{u_t^1 \in \mathcal{U}_t^1} \tilde{c}_t^1(\pi_t^1, a_t^2, u_t^1) \\ &+ \mathbb{E}[J_{t+1}^1(\Pi_{t+1}^1, A_{t+1}^2) \mid \pi_t^1, a_t^2, u_t^1], \quad t = 0, \ldots, T, \end{aligned} \tag{27}$$

and $J_{T+1}^1(\pi_{T+1}, a_{T+1}) := 0$ identically. The person-by-person optimal control law at time t is $u_t^{*1} = g_t^{*1}(\pi_t^1, a_t^2)$, i.e., the $\arg\inf$ in the RHS of (27), and the performance of the system is $\mathcal{J}^1(\boldsymbol{g}^{*1}) = \mathbb{E}[J_0^1(\Pi_0^1, A_0^2)]$.

However, we seek the best control strategy $\hat{\boldsymbol{g}}^{*1}$ for Problem 2 which takes the structural form $u_t^1 = \hat{g}_t^1(\hat{\pi}_t^1, a_t^2)$ for all $t = 0, \ldots, T$. Thus, we define the modified value functions

$$\begin{aligned} \hat{J}_t^1(\hat{\pi}_t^1, a_t^2) &:= \inf_{u_t^1 \in \mathcal{U}^1} \tilde{c}_t^1(\hat{\pi}_t^1, a_t^2, u_t^1) \\ &+ \mathbb{E}[\hat{J}_{t+1}^1(\hat{\Pi}_{t+1}^1, A_{t+1}^2) \mid \hat{\pi}_t^1, a_t^2, u_t^1], \quad t = 0, \ldots, T, \end{aligned} \tag{28}$$

and $\hat{J}_{t+1}^1(\hat{\pi}_t^1, a_t^2) := 0$ identically, where $\hat{\pi}_t^1 = \sigma(\pi_t^1)$. For a fixed $\boldsymbol{\psi}^2$, the best control law using the approximate information state at each t is $u_t^{*1} = \hat{g}_t^{*1}(\hat{\pi}_t^1, a_t^2)$, i.e., the $\arg\inf$ in the RHS of (28), and the performance of the system is $\mathcal{J}^1(\hat{\boldsymbol{g}}^{*1}) = \mathbb{E}[\hat{J}_0^1(\hat{\Pi}_0^1, A_0^2)]$. The *loss in person-by-person performance* which arises from using the approximate information state is measured by the difference $|\mathcal{J}^1(\boldsymbol{g}^{*1}) - \mathcal{J}^1(\hat{\boldsymbol{g}}^{*1})|$. Next, we present a result for this performance loss.

Lemma 10. *For any given prescription strategy $\boldsymbol{\psi}^2$,*

$$\lim_{n \to \infty} |\mathcal{J}^1(\boldsymbol{g}^{*1}) - \mathcal{J}^1(\hat{\boldsymbol{g}}^{*1})| = 0. \tag{29}$$

Proof. The proof follows directly from [20, Theorem 3]. □

Lemma 10 establishes the asymptotic convergence of the optimal performance by using the approximate information state towards the exact person-by-person optimal performance. Furthermore, it implies that for any desired upper bound on loss $\alpha_0 \in \mathbb{R}_{\geq 0}$, there exists a number $n \in \mathbb{N}$ and set $\mathcal{Q}_n$, such that $|\mathcal{J}^1(\boldsymbol{g}^{*1}) - \mathcal{J}^1(\hat{\boldsymbol{g}}^{*1})| < \alpha_0$. An explicit relationship between the upper bound α_0 and the upper bound on total variation distance, ϵ can be obtained using Theorem 9 and Proposition 46 of [21]. This is given by recursively defining

$$\alpha_t = 2(\epsilon \cdot ||c||_\infty + 3\epsilon \cdot ||\hat{J}^1_{t+1}||_\infty + 3\epsilon \cdot \hat{J}^1_L + \alpha_{t+1}), \quad (30)$$

where $||\hat{J}^1_{t+1}||_\infty := \sup_{\hat{\pi}^1_t, a^2_t} \hat{J}^1_{t+1}(\hat{\pi}^1_{t+1}, a^2_{t+1})$ and $\hat{J}^1_L$ is a finite upper bound on the Lipschitz constant of $\hat{J}^1_t$ for all $t = 0, \dots, T$. Note that an upper bound on the value of $\hat{J}_t$ exists for all $t = 0, \dots, T$ because cost is upper bounded. Furthermore, the Lipschitz continuity of $\hat{J}^1_t$ arises naturally from the fact that it is piece-wise linear and concave with respect to $\hat{\pi}^1_t$ for all $t = 0, \dots, T$ [22].

The maximum loss in person-by-person performance from using an approximate information state in $\mathcal{Q}_n$ is $||\alpha_0||_\infty := \sup_{\boldsymbol{\psi}^2} \alpha_0$. Furthermore, we define an approximate information state for agent 2 as $\hat{\Pi}^2_t := \mathbb{P}^{\boldsymbol{\psi}}(X^1_t, L^2_t, \hat{\Pi}^1_t \mid M^2_t, \Gamma^{1:2}_{0:t-1})$. In a manner similar to Lemma 4, we can show that at each $t = 0, \dots, T-1$, there exists a function $\hat{f}^2_t$ such that $\hat{\Pi}^2_{t+1} = \hat{f}^2_t(\hat{\Pi}^2_t, \Gamma^{1:2}_t, Z^2_{t+1})$. Thus, using the same sequence of arguments as Theorem 2, we conclude that if we restrict our attention to control strategies with the structural form $U^1_t = \hat{g}^1_t(\hat{\Pi}^1_t, \hat{\Pi}^2_t)$, and $U^2_t = \hat{g}^2_t(L^2_t, \hat{\Pi}^2_t)$ for all $t = 0, \dots, T$, the maximum loss in optimal performance in Problem 1, $|\mathcal{J}(\boldsymbol{g}^*) - \mathcal{J}(\hat{\boldsymbol{g}}^{*1}, \hat{\boldsymbol{g}}^{*1})|$, is also $||\alpha_0||_\infty$.

Remark 10. In this approximation technique, the set of feasible values of $\hat{\Pi}^1_t$, $\mathcal{Q}_n$, is finite and does not grow in size with time. Thus, $\hat{\Pi}^2_t$ is a simple probability distribution with a finite support, which, in turn, simplifies the implementation of our DP. However, it is still challenging to compute globally optimal prescription strategies for moderate and large values of the parameter $n \in \mathbb{N}$ because the number of possible prescriptions of agent 1, $|\mathcal{U}^1_t|^{|\mathcal{Q}_n|}$, grows exponentially with n. Instead, this approach may be utilized when only person-by-person optimal strategies are required.

VI. Conclusions

In this paper, we introduced a general model for decentralized control of two agents with nested accessible information. We derived structural forms for optimal control strategies with domains which do not grow in size with time and thus, can be derived using a DP decomposition. We also presented simplified optimal control strategies for systems with decoupled state and observation dynamics. Finally, we presented an approximate information state which can be used to derive approximately optimal control strategies with smaller domains. One potential direction for future research includes deriving more efficient approximate representations of the information state and prescriptions. Another important direction of future research is the development of approximate algorithms specialized to efficiently solve the DPs which arise in decentralized control.

References

[1] A. A. Malikopoulos, L. Beaver, and I. V. Chremos, "Optimal time trajectory and coordination for connected and automated vehicles," *Automatica*, vol. 125, p. 109469, 2021.

[2] A. Dave, I. V. Chremos, and A. A. Malikopoulos, "Social Media and Misleading Information in a Democracy: A Mechanism Design Approach," *IEEE Transactions on Automatic Control*, 2022 (in press).

[3] A. Mahajan, N. C. Martins, M. C. Rotkowitz, and S. Yüksel, "Information structures in optimal decentralized control," in *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, pp. 1291–1306, IEEE, 2012.

[4] P. R. Kumar and P. P. Varaiya, *Stochastic Systems: Estimation, Identification, and Adaptive Control*. Englewood Cliffs, NJ: Prentice-Hall, 1986.

[5] L. Lessard and A. Nayyar, "Structural results and explicit solution for two-player lqg systems on a finite time horizon," in *52nd IEEE Conference on Decision and Control*, pp. 6542–6549, IEEE, 2013.

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

[7] A. Mahajan, "Optimal decentralized control of coupled subsystems with control sharing," *IEEE Transactions on Automatic Control*, vol. 58, no. 9, pp. 2377–2382, 2013.

[8] A. Nayyar and D. Teneketzis, "On the structure of real-time encoding and decoding functions in a multiterminal communication system," *IEEE transactions on information theory*, vol. 57, no. 9, pp. 6196–6214, 2011.

[9] A. Dave and A. A. Malikopoulos, "Decentralized Stochastic Control in Partially Nested Information Structures," in *IFAC-PapersOnLine*, vol. 52, (Chicago, IL, USA), pp. 97–102, 2019.

[10] A. Dave and A. A. Malikopoulos, "A dynamic program for a team of two agents with nested information," in *60th IEEE Conference on Decision and Control (CDC)*, pp. 3768–3773, IEEE, 2021.

[11] Y. Xie, J. Dibangoye, and O. Buffet, "Optimally solving two-agent decentralized pomdps under one-sided information sharing," in *International Conference on Machine Learning*, pp. 10473–10482, PMLR, 2020.

[12] A. Dave and A. A. Malikopoulos, "Structural results for decentralized stochastic control with a word-of-mouth communication," in *2020 American Control Conference (ACC)*, pp. 2796–2801, IEEE, 2020.

[13] A. A. Malikopoulos, "On team decision problems with nonclassical information structures," *IEEE Transactions on Automatic Control*, 2022 arXiv:2101.10992.

[14] D. S. Bernstein, R. Givan, N. Immerman, and S. Zilberstein, "The complexity of decentralized control of markov decision processes," *Mathematics of operations research*, vol. 27, no. 4, pp. 819–840, 2002.

[15] S. M. Asghari, Y. Ouyang, and A. Nayyar, "Optimal local and remote controllers with unreliable uplink channels," *IEEE Transactions on Automatic Control*, vol. 64, no. 5, pp. 1816–1831, 2018.

[16] Y. Ouyang, S. M. Asghari, and A. Nayyar, "Optimal local and remote controllers with unreliable communication," in *2016 IEEE 55th Conference on Decision and Control (CDC)*, pp. 6024–6029, IEEE, 2016.

[17] A. Dave, N. Venkatesh, and A. A. Malikopoulos, "On decentralized minimax control with nested subsystems," *2022 American Control Conference (ACC)*, 2022 (to appear).

[18] A. Nayyar and D. Teneketzis, "Common knowledge and sequential team problems," *IEEE Transactions on Automatic Control*, vol. 64, no. 12, pp. 5108–5115, 2019.

[19] Y. A. Reznik, "An algorithm for quantization of discrete probability distributions," in *2011 Data Compression Conference*, pp. 333–342, IEEE, 2011.

[20] N. Saldi, S. Yüksel, and T. Linder, "Asymptotic optimality of finite model approximations for partially observed markov decision processes with discounted cost," *IEEE Transactions on Automatic Control*, vol. 65, no. 1, pp. 130–142, 2019.

[21] J. Subramanian, A. Sinha, R. Seraj, and A. Mahajan, "Approximate information state for approximate planning and reinforcement learning in partially observed systems," *Journal of Machine Learning Research*, vol. 23, no. 12, pp. 1–83, 2022.

[22] R. D. Smallwood and E. J. Sondik, "The optimal control of partially observable markov processes over a finite horizon," *Operations research*, vol. 21, no. 5, pp. 1071–1088, 1973.