2021 60th IEEE Conference on Decision and Control (CDC)
December 13-15, 2021. Austin, Texas

# A Dynamic Program for a Team of Two Agents with Nested Information

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

***Abstract*— In this paper, we investigate a sequential dynamic team problem consisting of two agents with a nested information structure. We use a combination of the person-by-person and prescription approach to derive structural results for optimal control strategies for the team. We then use these structural results to present a dynamic programming (DP) decomposition to derive the optimal control strategies for a finite time horizon. We show that our DP utilizes the nested information structure to simplify the computation of the optimal control laws for the team at the final time step.**

## I. Introduction

Team theory refers to problems where a team of agents seeks to cooperatively control a state and minimize a shared cost [1], with applications including connected and automated vehicles [2], social media platforms [3], and robot swarms [4]. A key aspect of these problems is the team's *information structure*, which describes the information available to each agent at any time. Various information structures are categorized as: (1) *Classical:* Each agent receives the same information and has perfect recall [5]. (2) *Quasi-classical:* If agent 1 can affect the information of agent 2, the information available to agent 1 is *also* available to agent 2 [6]–[10]. (3) *Non-classical:* All other decentralized information structures are called non-classical [11]–[14].

In this paper, we analyze a dynamic team of two agents with a *nested information structure*. In the nested information structure, agent 2 shares her information with agent 1 at each instance of time, but does not receive any information from agent 1. Both agents collectively control and partially observe a shared state. Thus, this is a non-classical information structure. The nested information structure is commonly found in applications with real time communication problems [15], vehicle platoons [16], and hierarchical control problems [17]. A team of two agents with a quasi-classical, partially nested information structure, is also special case of the nested team when agent 2 can affect the state of agent 1, for example [8]. The two agent team is of interest because most insights into the structure of optimal control strategies in a team of two agents can be extended teams of many agents with a similar information structure. Such partially nested teams are well understood in the literature for linear dynamics, quadratic costs, and Gaussian noise (LQG) [6]–[8] and for nonlinear dynamics with complete state observation [17]. Several dynamic program (DP) decompositions have been reported in the literature for decentralized teams [18], [19], which are reviewed in detail in [14], [20]. Furthermore, a DP that can be applied in team problems with nested information structures is the one proposed in *the common information approach* [11], [18], which introduces a coordinator who selects prescription functions for each agent.

Our main contribution in this paper is to present structural results for optimal control strategies in the nested information structure which cannot be derived solely using the common information approach. Our analysis uses a combination of *the person-by-person approach* [6]–[8], [21] and *the prescription approach* [13]. Similar techniques have been used in conjunction with linear dynamics [8], where the dynamics ensure that the optimal control strategies depend only on the expected value of certain random variables. This ensures the tractability of the eventual DP. For nonlinear dynamics, similar techniques have been used for real time communication [15] and control sharing information structures [22] by assuming specific dynamics. However, when agents imperfectly observe their states, the optimal control strategies derived in these papers are functions of non-parametric probability distribution with a continuous support. Thus, it is challenging to actually implement these optimal strategies. In contrast, we derive optimal control strategies without assuming specific dynamics and our strategies only require tracking probability distributions with an atomic support. However, our structural results yield strategies whose domain grows in size with time and thus, can only be applied to finite time horizons. Despite this, we believe that our results may be useful in the search for approximately optimal control strategies [23] that may be time invariant. We also present a DP based on our structural results. Our DP deviates from other DPs in the literature for the final time step, where we utilize the the nested information structure to improve the computational efficiency of optimal control strategies.

The remainder of the paper proceeds as follows. In Section II, we provide the problem formulation. In Section III, we analyze a team of two agents using the person-by-person and prescription approaches, and derive structural results for optimal control strategies. In Section IV, we present a DP to derive the optimal control strategies. Finally, in Section V, we present concluding remarks and discuss ongoing work.

## II. Problem Formulation

We consider a team of two agents who select actions over $T \in \mathbb{N}$ discrete time steps. At any time $t = 0, \dots, T$, the state of the team is denoted by the random variable $X_t$ that takes values in a finite set of feasible states $\mathcal{X}_t$. The control action of agent $k \in \{1, 2\}$ at each time $t$ is denoted by the random variable $U_t^k$ that takes values in a finite set of

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; andreas@udel.edu).

feasible actions $\mathcal{U}_t^k$. Let $U_t^{1:2} = (U_t^1, U_t^2)$. Starting with the initial state $X_0$ at $t = 0$, the system evolves as

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

where the random variable $W_t$ denotes an uncontrolled disturbance to the state at time $t$ and takes values in a finite set of feasible disturbances $\mathcal{W}_t$. At each time $t = 0, \dots, T$, each agent $k \in \{1, 2\}$ partially observes the state as a random variable $Y_t^k$ that takes values in a finite set $\mathcal{Y}_t^k$. The observation $Y_t^k$ is given by

$$Y_t^k = h_t^k(X_t, V_t^k), \quad t = 0, \dots, T, \tag{2}$$

where the random variable $V_t^k$ denotes the measurement noise that takes values in a finite set of feasible noises $\mathcal{V}_t^k$. The external disturbances $\{W_t : t = 0, \dots, T\}$, noises in measurement $\{V_t^1, V_t^2 : t = 0, \dots, T\}$, 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. This ensures that the state $X_t$ evolves as a controlled Markov chain at each $t = 0, \dots, T$ [5]. Next, we define the *memory* of each agent $k \in \{1, 2\}$ at each time $t$, which is a collection of all the data received by them. The memories of both agents determine the team's *information structure*.

**Definition 1.** The memory of agent $k \in \{1, 2\}$ at each time $t = 0, \dots, T$ is a set of random variables $M_t^k$ that takes values in a finite collection of sets $\mathcal{M}_t^k$.

Let $U_{0:t}^k = (U_0^k, \dots, U_t^k)$, $k \in \{1, 2\}$. The memories of the two agents in the team at any time $t = 0, \dots, T$ are given by

$$M_t^2 := \{Y_{0:t}^2, U_{0:t-1}^2\}, \tag{3}$$
$$M_t^1 := \{Y_{0:t}^1, U_{0:t-1}^1, Y_{0:t}^2, U_{0:t-1}^2\}. \tag{4}$$

In (3)-(4), we consider that each agent updates her memory before generating her action at each time $t$. Note that the memories satisfy the following properties for all $t = 0, \dots, T$: (1) *causality:* $M_t^k \subseteq \{Y_{0:t}^{1:2}, U_{0:t-1}^{1:2}\}$, $k \in \{1, 2\}$; (2) *perfect recall:* $M_t^k \subseteq M_{t+1}^k$, $k \in \{1, 2\}$; and (3) *nested information structure:* $M_t^2 \subseteq M_t^1$.

**Remark 1.** Causality and perfect recall are very general properties used to derive DPs for both centralized [5] and decentralized [20] teams. The nested information structure is unique to our team.

The new information of any agent $k$ at each $t$ is the set $Z_t^k := M_t^k \setminus M_{t-1}^k$ that takes values in a finite collection of sets $\mathcal{Z}_t^k$. Thus,

$$Z_t^1 = \{Y_t^1, U_{t-1}^1, Y_t^2, U_{t-1}^2\}, \quad t = 0, \dots, T, \tag{5}$$
$$Z_t^2 = \{Y_t^2, U_{t-1}^2\}, \quad t = 0, \dots, T, \tag{6}$$

which implies that

$$Z_t^2 \subset Z_t^1, \quad t = 0, \dots, T. \tag{7}$$

Each agent $k \in \{1, 2\}$ at each $t = 0, \dots, T$ selects an action $U_t^k$ as a function of her memory $M_t^k$. Thus,

$$U_t^k = g_t^k(M_t^k), \quad t = 0, \dots, T, \tag{8}$$

where $g_t^k$ is the *control law* of agent $k$ at time $t$. The *control strategy* of each agent $k$ is $\boldsymbol{g}^k := (g_0^k, \dots, g_T^k)$ and the *strategy profile* of the team is $\boldsymbol{g} := (\boldsymbol{g}^1, \boldsymbol{g}^2)$. The set of all feasible strategy profiles is $\mathcal{G}$. After each $k \in \{1, 2\}$ selects action $U_t^k$ at time $t$, the team incurs a cost $c_t(X_t, U_t^1, U_t^2) \in \mathbb{R}_{\geq 0}$. Then, the performance criterion for the system is

$$\mathcal{J}(\boldsymbol{g}) = \mathbb{E}^{\boldsymbol{g}}\left[\sum_{t=0}^{T} c_t\big(X_t, U_t^1, U_t^2\big)\right], \tag{9}$$

where the expectation is with respect to the joint distribution on all random variables, and $U_t^k$ is given by (8) for each agent $k \in \{1, 2\}$, at each time $t = 0, \dots, T$. Then, we can state the optimization problem for the team as follows.

**Problem 1.** The optimization problem is $\inf_{\boldsymbol{g} \in \mathcal{G}} \mathcal{J}(\boldsymbol{g})$, given the probability distributions of the primitive random variables $\{X_0, W_t, V_t^1, V_t^2 : t = 0, \dots, T\}$, and the functions $\big\{c_t, f_t, h_t^{1:2} : t = 0, \dots, T\big\}$.

Our aim is to develop a DP that can tractably derive an optimal strategy profile $\boldsymbol{g}^* \in \mathcal{G}$ for Problem 1, such that $\mathcal{J}(\boldsymbol{g}^*) \leq \mathcal{J}(\boldsymbol{g})$, for all $\boldsymbol{g} \in \mathcal{G}$.

## III. Analysis

### A. The Person-by-Person Approach

In this subsection, we present a structural result for the optimal control strategy of agent 1 using the person-by-person approach. This will help us derive our DP in Section IV. We first fix a control strategy $\boldsymbol{g}^2$ for agent 2, such that

$$U_t^2 = g_t^2(M_t^2), \quad t = 0, \dots, T. \tag{10}$$

In this approach, given the strategy $\boldsymbol{g}^2$ of agent 2, we set up a centralized problem from the perspective of agent 1. Since $M_t^2 \subseteq M_t^1$ at each time $t = 0, \dots, T$, given the control strategy $\boldsymbol{g}^2$, agent 1 can derive the action $U_t^2$ using (10). Then, we can define a new state for agent 1 as

$$S_t^1 := \{X_t, M_t^2\}, \quad t = 0, \dots, T, \tag{11}$$

that takes values in a finite collection of sets $\mathcal{S}_t^1$ at any time $t$. Next, we show that the new state is sufficient for input-output mapping.

**Lemma 1.** *Let $\boldsymbol{g}^2$ be a given control strategy of agent 2. At each time $t = 0, \dots, T$, the state $S_t^1 \in \mathcal{S}_t^1$ is sufficient for input-output mapping by the following properties [24]:*

*1) There exist functions $\hat{f}_t^1(\cdot)$ and $\hat{h}_t^1(\cdot)$ for all $t = 0, \dots, T-1$, such that*

$$S_{t+1}^1 = \hat{f}_t^1(S_t^1, U_t^1, W_t, V_{t+1}^{1:2}), \tag{12}$$
$$Z_{t+1}^1 = \hat{h}_t^1(S_t^1, U_t^1, W_t, V_{t+1}^{1:2}). \tag{13}$$

*2) There exist functions $\hat{c}_t^1(\cdot)$, such that*

$$c_t(X_t, U_t^1, U_t^2) = \hat{c}_t^1(S_t^1, U_t^1), \quad t = 0, \dots, T. \tag{14}$$

*Proof.* To prove these results, we expand the LHS in each of (12)-(14) by substituting appropriate relations from the system dynamics (1), (2), the definitions (5), (11), and (10), for each $t = 0, \dots, T$. Thus, we can rewrite the LHS in terms of the variables in the RHS and construct appropriate functions $\hat{f}_t^1(\cdot)$, $\hat{h}_t^1(\cdot)$, and $\hat{c}_t^1(\cdot)$ for each time $t$. □

Given the strategy $\boldsymbol{g}^2$, Lemma 1 leads to a centralized problem for agent 1, with state $S_t^1$, control action $U_t^1$, observation $Z_{t+1}^1$, and cost $\hat{c}_t^1(S_t^1, U_t^1)$ at each time $t = 0, \dots, T$. The performance criterion is solely a function of the control strategy $\boldsymbol{g}^1$, as $\mathcal{J}^1(\boldsymbol{g}^1) = \mathbb{E}^{\boldsymbol{g}}\left[\sum_{t=0}^T \hat{c}_t^1(S_t^1, U_t^1)\right]$, where the expectation is with respect to the joint probability distribution on all random variables and $\boldsymbol{g} = (\boldsymbol{g}^1, \boldsymbol{g}^2)$.

**Problem 2.** The problem for agent 1 is $\inf_{\boldsymbol{g}^1} \mathcal{J}^1(\boldsymbol{g}^1)$, given the control strategy $\boldsymbol{g}^2$, the probability distributions of the primitive random variables $\{X_0, W_t, V_t^1, V_t^2 : t = 0, \dots, T\}$, and the functions $\{c_t, f_t, h_t^{1:2} : t = 0, \dots, T\}$.

In Problem 2, at each time $t$, the component $M_t^2$ of the state $S_t^1$ is observed by agent 1. However, the component $X_t$ of state $S_t^1$ must be inferred by agent 1 using her memory $M_t^1$. For such a problem, it is known [5, page 79] that agent 1 can estimate $X_t$ using the probability distribution

$$\Pi_t^1 := \mathbb{P}^{\boldsymbol{g}}\big(X_t \mid M_t^1\big), \quad t = 0, \dots, T, \tag{15}$$

that takes values in the set of feasible distributions $\mathcal{P}_t^1 := \Delta(\mathcal{X}_t)$ at each time $t$. The distribution $\Pi_t^1$ is called an *information state* for agent 1 and yields the following structural result for the control strategy of agent 1 in Problem 2.

**Theorem 1.** *Let $\boldsymbol{g}^2$ be a given control strategy for agent 2. Then, the optimal control strategy $\boldsymbol{g}^{*1}$ of agent 1 in Problem 2 has the structural form*

$$U_t^1 = g_t^{*1}(M_t^2, \Pi_t^1), \quad t = 0, \dots, T. \tag{16}$$

*Proof.* This result follows from standard arguments for partially observed Markov decision processes [5, page 79]. □

Note that every optimal strategy profile $\boldsymbol{g}^* = (\boldsymbol{g}^{*1}, \boldsymbol{g}^{*2})$ for Problem 1, must be a solution of Problem 2 by fixing $\boldsymbol{g}^{*2}$ for agent 2 and selecting the control strategy of agent 1 as $\boldsymbol{g}^{*1}$ [6]. Thus, every optimal profile $\boldsymbol{g}^*$ for Problem 1 also satisfies Theorem 1. Then, in Problem 1, we can restrict our attention to strategy profiles $\boldsymbol{g} \in \mathcal{G}$ with the structural form

$$U_t^1 = g_t^1(M_t^2, \Pi_t^1), \tag{17}$$
$$U_t^2 = g_t^2(M_t^2), \tag{18}$$

at each time $t = 0, \dots, T$. To this end, we denote the set of feasible strategy profiles consistent with (17)-(18) by $\mathcal{G}'$.

### *B. The Prescription Approach*

In this subsection, we consider Problem 1 with the restriction $\boldsymbol{g} \in \mathcal{G}'$. Any strategy profile $\boldsymbol{g} \in \mathcal{G}'$ for the team is accessible to both agents. However, at any time $t = 0, \dots, T$, agent 2 cannot generate the action $U_t^1$ using (17), because she can only access the memory $M_t^2$ and not the information state $\Pi_t^1 \in \mathcal{P}_t^1$, which is a function of the memory $M_t^1$. Instead, agent 2 considers that the action $U_t^1$ is generated in two stages at each time $t$: (1) agent 1 generates a function using only $M_t^2$, and (2) this function takes as an input the information state $\Pi_t^1$ to generate the action $U_t^1$. We call this function a *prescription* of agent 2 for agent 1 at time $t$.

**Definition 2.** A *prescription* of agent 2 for agent 1 at any time $t = 0, \dots, T$ is a function $\Gamma_t^{[2,1]} : \mathcal{P}_t^1 \to \mathcal{U}_t^1$ that takes values in a finite set of feasible functions $\mathcal{F}_t^{[2,1]}$.

The prescription $\Gamma_t^{[2,1]}$ is generated as

$$\Gamma_t^{[2,1]} = \psi_t^{[2,1]}(M_t^2), \quad t = 0, \dots, T, \tag{19}$$

where $\psi_t^{[2,1]} : \mathcal{M}_t^2 \to \mathcal{F}_t^{[2,1]}$ is called the prescription law of agent 2 for agent 1 at time $t$. We call $\boldsymbol{\psi}^2 := \big(\psi_t^{[2,1]} : t = 0, \dots, T\big)$ the prescription strategy of agent 2, and denote the set of feasible prescription strategies by $\Psi^2$. Next, Lemmas 2 and 3 show that any control action $U_t^1$ can be equivalently generated using either a control strategy $\boldsymbol{g} \in \mathcal{G}'$ or an appropriate prescription strategy $\boldsymbol{\psi}^2$.

**Lemma 2.** *For any given control strategy $\boldsymbol{g} \in \mathcal{G}'$, we can construct a prescription strategy $\boldsymbol{\psi}^2 \in \Psi^2$ such that*

$$\Gamma_t^{[2,1]}\big(\Pi_t^1\big) = g_t^1(M_t^2, \Pi_t^1) = U_t^1, \quad t = 0, \dots, T. \tag{20}$$

*Proof.* For any control law $g_t^1$ that generates $U_t^1$ at time $t$ using (17), we can construct a prescription law $\psi_t^{[2,1]} : \mathcal{M}_t^2 \to \mathscr{F}_t^{[2,1]}$ as $\Gamma_t^{[2,1]}(\Pi_t^1) = \psi_t^{[2,1]}(M_t^2)(\Pi_t^1) := g_t^1(M_t^2, \Pi_t^1) = U_t^1$, for all $t = 0, \dots, T$. □

**Lemma 3.** *For any given prescription strategy $\boldsymbol{\psi}^2 \in \Psi^2$, we can construct a control strategy $\boldsymbol{g} \in \mathcal{G}'$ such that*

$$g_t^1(M_t^2, \Pi_t^1) = \Gamma_t^{[2,1]}(\Pi_t^1) = U_t^1, \quad t = 0, \dots, T. \tag{21}$$

*Proof.* For any prescription strategy $\boldsymbol{\psi}^2$, we construct a control strategy $\boldsymbol{g}$ such that $g_t^1(M_t^2, \Pi_t^1) := \psi_t^{[2,1]}(M_t^2)(\Pi_t^1) = \Gamma_t^{[2,1]}(\Pi_t^1) = U_t^1$, for all $t = 0, \dots, T$. □

Lemmas 2 and 3 imply that every control action $U_t^1$ of an agent 1 generated through a control strategy $\boldsymbol{g}^1$ can also be generated through an appropriate prescription strategy $\boldsymbol{\psi}^2$ and vice versa. We always choose control and prescription strategies that satisfy (20) and (21). Thus, at each time $t$,

$$\Pi_t^1 = \mathbb{P}^{\boldsymbol{g}}(X_t \mid M_t^1) = \mathbb{P}^{(\boldsymbol{g}, \boldsymbol{\psi}^2)}(X_t \mid M_t^1, \Gamma_{0:t-1}^{[2,1]}), \tag{22}$$

where we use Lemma 2 to construct $\boldsymbol{\psi}^2$ given $\boldsymbol{g}$, and we can add the history of prescriptions $\Gamma_{0:t-1}^{[2,1]}$ to the conditioning because they are simply functions of $M_t^2 \subseteq M_t^1$ and $\boldsymbol{\psi}^2$. Using similar arguments, it holds that $\mathcal{J}(\boldsymbol{g}) = \mathbb{E}^{(\boldsymbol{g}, \boldsymbol{\psi}^2)}\left[\sum_{t=0}^T c_t(X_t, U_t^{1:2})\right]$, where $U_t^1$ is equivalently generated using either $g_t^1(M_t^2, \Pi_t^1)$ or $\Gamma_t^{[2,1]}(\Pi_t^1)$.

### *C. A New State for Agent 2*

In this subsection, we define a state sufficient for input-output mapping for agent 2. We first define the information

accessible to agent 1 but *inaccessible* to agent 2 at each time $t = 0, \dots, T$ as the set of random variables

$$L_t^{[1,2]} := M_t^1 \setminus M_t^2, \tag{23}$$

that takes values in a finite collection of sets $\mathcal{L}_t^{[1,2]}$. For all $t$, we define an information state for agent 2 as the distribution

$$\Pi_t^2 := \mathbb{P}^{(\boldsymbol{g},\boldsymbol{\psi}^2)}\big(X_t, L_t^{[1,2]} \mid M_t^2, \Gamma_{0:t-1}^{[2,1]}\big), \tag{24}$$

that takes values in the set of feasible distributions $\mathcal{P}_t^2 := \Delta(\mathcal{X}_t \times \mathcal{L}_t^{[1,2]})$. Next, we show that we can write the information state $\Pi_t^1$ of agent 1 in terms of $\Pi_t^2$ at each $t = 0, \dots, T$.

**Lemma 4.** *At any time $t = 0, \dots, T$, for the pair of probability distributions $\Pi_t^1$ and $\Pi_t^2$, we can construct a function $e_t : \mathcal{P}_t^2 \times \mathcal{L}_t^{[1,2]} \to \mathcal{P}_t^1$, such that*

$$\Pi_t^1 = e_t\big(\Pi_t^2, L_t^{[1,2]}\big). \tag{25}$$

*Proof.* The proof is omitted due to space constraints, but can be found in our online preprint [25, Appendix A]. □

Thus, at each time $t$, we can equivalently write the control action as $U_t^1 = \Gamma_t^{[2,1]}(\Pi_t^1) = \Gamma_t^{[2,1]}\big(e_t(\Pi_t^2, L_t^{[1,2]})\big)$. Next, we construct a new state for agent 2 as

$$S_t^2 := \big\{X_t, L_t^{[1,2]}, \Pi_t^2\big\}, \quad t = 0, \dots, T, \tag{26}$$

that takes values in the finite collection of sets $\mathcal{S}_t^2$. Our goal is to set up an equivalent centralized control problem for agent 2 with the state $S_t^2$ and control action $(\Gamma_t^{[2,1]}, U_t^2)$ at each time $t$. However, we require some interim results before we can prove that the state $S_t^2$ is sufficient for input-output mapping. Next, we show that the information states $\Pi_t^2$ and $\Pi_t^1$ at all $t$ are independent from the strategies $(\boldsymbol{g}, \boldsymbol{\psi}^2)$.

**Lemma 5.** *At each time $t = 0, \dots, T-1$, there exists a function $\tilde{f}_t^2(\cdot)$ independent from $(\boldsymbol{g}, \boldsymbol{\psi}^2)$, such that*

$$\Pi_{t+1}^2 = \tilde{f}_t^2(\Pi_t^2, \Gamma_t^{[2,1]}, U_t^2, Z_{t+1}^2), \tag{27}$$

*and subsequently, for any Borel subset $P^2 \subset \mathcal{P}_{t+1}^2$,*

$$\mathbb{P}(\Pi_{t+1}^2 \in P^2 |\ M_t^2, \Gamma_{0:t}^{[2,1]}) = \mathbb{P}(\Pi_{t+1}^2 \in P^2 |\ \Pi_t^2, U_t^2, \Gamma_t^{[2,1]}).$$

*Proof.* The proof is omitted due to space constraints, but can be found in our online preprint [25, Appendix B]. □

**Lemma 6.** *At each time $t = 0, \dots, T-1$, there exists a function $\tilde{f}_t^1(\cdot)$ independent from $(\boldsymbol{g}, \boldsymbol{\psi}^2)$, such that*

$$\Pi_{t+1}^1 = \tilde{f}_t^1(\Pi_t^1, U_t^1, U_t^2, Z_{t+1}^1), \tag{28}$$

*and subsequently, for any Borel subset $P^1 \subset \mathcal{P}_{t+1}^1$,*

$$\mathbb{P}(\Pi_{t+1}^1 \in P^1 |\ M_t^1, U_t^2) = \mathbb{P}(\Pi_{t+1}^1 \in P^1 |\ \Pi_t^1, U_t^1, U_t^2).$$

*Proof.* The proof is similar to Lemma 5 and is omitted. □

**Lemma 7.** *At each time $t = 0, \dots, T$, there exists a function $\tilde{c}_t^k(\cdot)$ for each $k \in \{1, 2\}$ such that*

$$\mathbb{E}^{\boldsymbol{g}}[c_t(X_t, U_t^{1:2}) | M_t^1, U_t^{1:2}] = \tilde{c}_t^1(\Pi_t^1, U_t^{1:2}), \tag{29}$$

$$\mathbb{E}^{\boldsymbol{g}}[\tilde{c}_t^1(\Pi_t^1, U_t^{1:2}) | M_t^2, \Gamma_t^{[2,1]}, U_t^2] = \tilde{c}_t^2(\Pi_t^2, \Gamma_t^{[2,1]}, U_t^2). \tag{30}$$

*Proof.* We first prove (29). Let $m_t^1$, $u_t^{1:2}$, and $\pi_t^1$ be the realizations of the random variables $M_t^1$, $U_t^{1:2}$, and the conditional distribution $\Pi_t^1$ at each time $t = 0, \dots, T$. Then, we expand the expectation as $\mathbb{E}^{\boldsymbol{g}}[c_t(X_t, U_t^{1:2}) \mid m_t^1, u_t^{1:2}] = \sum_{x_t} c_t(x_t, u_t^{1:2}) \cdot \mathbb{P}^{\boldsymbol{g}}(X_t = x_t \mid m_t^1, u_t^{1:2}) = \sum_{x_t} c_t(x_t, u_t^{1:2}) \cdot \pi_t^1(x_t) =: \tilde{c}_t^1(\pi_t^1, u_t^{1:2})$, where we can drop the control actions $u_t^{1:2}$ from the conditioning because they known given the strategy $\boldsymbol{g}$ and $m_t^1$. We prove (30) using the same arguments as above. □

Next, we prove that the state $S_t^2$ is sufficient for input-output mapping from the perspective of agent 2.

**Lemma 8.** *At each time $t$, the state $S_t^2 \in \mathcal{S}_t^2$ satisfies the following properties stated by Witsenhausen [24]:*

*1) There exist functions $\hat{f}_t^2(\cdot)$ and $\hat{h}_t^2(\cdot)$ for all $t = 0, \dots, T-1$, such that*

$$S_{t+1}^2 = \hat{f}_t^2(S_t^2, \Gamma_t^{[2,1]}, U_t^2, W_t, V_{t+1}^{1:2}), \tag{31}$$

$$Z_{t+1}^2 = \hat{h}_{t+1}^2(S_t^2, \Gamma_t^{[2,1]}, U_t^2, W_t, V_t^{1:2}). \tag{32}$$

*2) There exist functions $\hat{c}_t^2(\cdot)$, such that*

$$c_t(X_t, U_t^{1:2}) = \hat{c}_t^2(S_t^2, \Gamma_t^{[2,1]}, U_t^2), \quad t = 0, \dots, T. \tag{33}$$

*Proof.* To prove these results, we expand the LHS in each of (31)-(33) to write them in terms of the variables in the RHS, and construct appropriate functions $\hat{f}_t^1(\cdot)$, $\hat{h}_t^1(\cdot)$, and $\hat{c}_t^1(\cdot)$ for each time $t$. □

Lemma 8 allows us to construct a centralized stochastic control problem for agent 2 with state $S_t^2$ that evolves using (31), control action $(\Gamma_t^{[2,1]}, U_t^2)$, observation $Z_t^2$ given by (32), and cost $\hat{c}_t^2(S_t^2, \Gamma_t^{[2,1]}, U_t^2)$ at each time $t = 0, \dots, T$. Furthermore, the performance criterion can be written as a function of the prescription strategy $\boldsymbol{\psi}^2$ and control strategy $\boldsymbol{g}^2$ as $\mathcal{J}^2(\boldsymbol{\psi}^2, \boldsymbol{g}^2) = \mathbb{E}^{\boldsymbol{g}}\Big[\sum_{t=0}^T \hat{c}_t^2(S_t^2, \Gamma_t^{[2,1]}, U_t^2)\Big]$.

**Problem 3.** The problem for agent 2 is $\inf_{\boldsymbol{\psi}^2, \boldsymbol{g}^2} \mathcal{J}^1(\boldsymbol{\psi}^2, \boldsymbol{g}^2)$, given the probability distributions of the primitive random variables $\{X_0, W_t, V_t^{1:2} : t = 0, \dots, T\}$, and the functions $\big\{\hat{c}_t^2, \hat{f}_t^2, \hat{h}_t^2 : t = 0, \dots, T\big\}$.

In Problem 3, at each time $t$, the component $\Pi_t^2$ of the state $S_t^2$ is completely observed by agent 2. Furthermore, the unobserved component $\big\{X_t, L_t^{[1,2]}\big\}$ can be estimated by agent 2 using the probability distribution $\Pi_t^2$. This yields the following structural result for agent 2 in Problem 3.

**Theorem 2.** *For agent 2 in Problem 3, without loss of optimality, we can restrict attention to prescription strategies $\boldsymbol{\psi}^{*2}$ and control strategies $\boldsymbol{g}^{*2}$ with the structural form*

$$\Gamma_t^{[2,1]} = \psi_t^{*[2,1]}\big(\Pi_t^2\big), \quad t = 0, \dots, T, \tag{34}$$

$$U_t^2 = g_t^{*2}\big(\Pi_t^2\big), \quad t = 0, \dots, T. \tag{35}$$

*Proof.* This result follows from standard arguments for partially observed Markov decision processes [5, page 79]. □

Recall that using Lemmas 2 and 3, given a prescription strategy $\boldsymbol{\psi}^{*2}$ of the form in Theorem 2, we can construct a corresponding control strategy $\boldsymbol{g}^{*1}$ for agent 1 as

$g_t^{*1}(\Pi_t^{1:2}) := \psi_t^{*[2,1]}(\Pi_t^2)(\Pi_t^1)$, for all $t = 0, \dots, T$. Then, $g_t^{*1}$ and $\psi_t^{*[2,1]}$ yield the same control action $U_t^1$ at each time $t = 0, \dots, T$. Thus, we can derive an optimal team strategy $\boldsymbol{g}^*$ for Problem 1 with the structural form

$$U_t^1 = g_t^{*1}(\Pi_t^1, \Pi_t^2), \quad t = 0, \dots, T, \tag{36}$$

$$U_t^2 = g_t^{*2}(\Pi_t^2), \quad t = 0, \dots, T. \tag{37}$$

To this end, we denote the set of feasible team strategies consistent with (36) and (37) by $\mathcal{G}''$.

## IV. The Dynamic Program

In this section, we present a DP to solve Problem 1 using the information states. Recall that at each $t = 0, \dots, T$, the memory $M_t^2$ and subsequently, the information state $\Pi_t^2$ are available to agent 1. We extend the memory of agent 1 at the final time step $T$ to also include $U_T^2$. Any team strategy $\boldsymbol{g} \in \mathcal{G}''$ can be implemented using the extended memory $\{\Pi_T^1, \Pi_T^2, U_T^2\}$ for agent 1, by discarding $U_T^2$. Next, we show that the every strategy using the extended memory can also be implemented using only $\{\Pi_T^1, \Pi_T^2\}$.

**Lemma 9.** *Let $\bar{g}_T^1 : \mathcal{P}_T^1 \times \mathcal{P}_T^2 \times \mathcal{U}_T^2 \to \mathcal{U}_T^1$ be an extended control law for time $T$. Then, we can construct a control law $g_T^1 : \mathcal{P}_T^1 \times \mathcal{P}_T^2 \to \mathcal{U}_T^1$ such that*

$$U_T^1 = \bar{g}_T^1(\Pi_T^1, \Pi_T^2, U_T^2) = g_T^1(\Pi_T^1, \Pi_T^2). \tag{38}$$

*Proof.* The proof follows by substituting the relation $U_T^2 = g_T^2(\Pi_T^2)$ into the extended control law, and constructing $g_T^1(\cdot)$ as $g_T^1(\Pi_T^1, \Pi_T^2) := \bar{g}_T^1\big(\Pi_T^1, \Pi_T^2, g_T^2(\Pi_T^2)\big)$. □

Lemma 9 establishes that we can equivalently select either $\bar{g}_T^1$ or $g_T^1$ at time $T$, because they yield the same control action $U_T^1$. To this end, we simply denote the control law of agent 1 at time $T$ by $g_T^1$, even with the extended memory.

### A. The Value Functions

In this subsection, we construct the value functions and corresponding control laws for our DP. Let $u_t^k$ and $\pi_t^k$ be the realizations of the random variable $U_t^k$ and information state $\Pi_t^k$ for each $k \in \{1, 2\}$, for each $t = 0, \dots, T$. We recursively define two value functions at time $T$ as

$$J_T^1(\pi_T^1, u_T^2) := \inf_{u_T^1 \in \mathcal{U}_T^1} \tilde{c}_T^1(\pi_T^1, u_T^{1:2}), \tag{39}$$

$$J_T^2(\pi_T^2) := \inf_{u_T^2 \in \mathcal{U}_T^2} \mathbb{E}^{(\boldsymbol{g}, \boldsymbol{\psi}^2)}\big[J_T^1(\Pi_T^1, u_T^2) \mid \pi_T^2, u_T^2\big]. \tag{40}$$

The control law for agent 1 at time $T$ is $u_T^{*1} = g_T^{*1}(\pi_T^1, u_T^2)$, i.e., the $\arg\inf$ in the RHS of (39). The control law for agent 2 is $u_T^{*2} = g_T^{*2}(\pi_T^2)$, i.e., the $\arg\inf$ in the RHS of (40).

Next, at each time $t = T-1, \dots, 0$, we recursively define

$$J_t(\pi_t^2) := \inf_{u_t^2 \in \mathcal{U}_t^2, \gamma_t^{[2,1]} \in \mathcal{F}_t^{[2,1]}} \tilde{c}_t^2\big(\pi_t^2, \gamma_t^{[2,1]}, u_t^1\big) + \mathbb{E}^{(\boldsymbol{g}, \boldsymbol{\psi}^2)}\Big[J_{t+1}\big(\Pi_{t+1}^2\big) \mid \pi_t^2, \gamma_t^{[2,1]}, u_t^2\Big], \tag{41}$$

where at time $T-1$, by convention $J_T(\Pi_T^2) = J_T^2(\Pi_T^2)$. The prescription law at time $t$ is $\gamma_t^{*[2,1]} = \psi_t^{*[2,1]}(\pi_t^2)$ and the control law of agent 2 is $u_t^{*2} = g_t^{*2}(\pi_t^2)$, i.e., the $\arg\inf$ in the RHS of (41). The value functions (39)-(41) and corresponding control laws form a DP for the team.

### B. Optimality of the Dynamic Program

In this subsection, we prove the optimality of our DP, starting with time $T$. Let $\boldsymbol{g}_t = (g_t^1, g_t^2)$ and $\boldsymbol{g}_{0:t} = (\boldsymbol{g}_0, \dots, \boldsymbol{g}_t)$. Furthermore, let $\mathcal{J}_t(\boldsymbol{g}) := \mathbb{E}^{(\boldsymbol{g}, \boldsymbol{\psi}^2)}\left[\sum_{\ell=t}^T c_\ell(X_\ell, U_\ell^{1:2})\right]$. Next, we show that the control law $g_T^{*1}$ is optimal for agent 1 at time $T$.

**Lemma 10.** *1) The value function $J_T^1$ in* (39) *is such that*

$$\mathcal{J}_T(\boldsymbol{g}) \geq \mathbb{E}^{(\boldsymbol{g}, \boldsymbol{\psi}^2)}[J_T^1(\Pi_T^1, U_T^2)], \quad \forall \boldsymbol{g} \in \mathcal{G}''. \tag{42}$$

*2) The corresponding control law $g_T^{*1}$ is such that*

$$\mathcal{J}_T(\boldsymbol{g}_{0:T-1}, g_T^{*1}, g_T^2) = \mathbb{E}^{(\boldsymbol{g}, \boldsymbol{\psi}^2)}[J_T^1(\Pi_T^1, U_T^2)]. \tag{43}$$

*Proof.* 1) Using the extended memory of agent 1 at time $T$, $U_T^1 = g_T^1(\Pi_T^{1:2}, U_T^2)$. Using Lemma 7, we write that

$$\mathcal{J}_T(\boldsymbol{g}) = \mathbb{E}^{(\boldsymbol{g}, \boldsymbol{\psi}^2)}\big[\tilde{c}_T^1(\Pi_T^1, U_T^{1:2})\big] \geq \mathbb{E}^{(\boldsymbol{g}, \boldsymbol{\psi}^2)}\big[J_T^1(\Pi_T^1, U_T^2)\big],$$

where, in the inequality, we used the definition of $J_T^1$ in (39).

2) We substitute $U_T^1 = g_T^{*1}(\Pi_T^1, U_T^2)$ in the expansion of $\mathcal{J}_T$, to write that

$$\begin{aligned}\mathcal{J}_T(\boldsymbol{g}_{0:T-1}, g_T^{*1}, g_T^2) &= \mathbb{E}^{(\boldsymbol{g}, \boldsymbol{\psi}^2)}\big[\tilde{c}_T^1(\Pi_T^1, g_T^{*1}(\Pi_T^1, U_T^2), U_T^2)\big] \\ &= \mathbb{E}^{(\boldsymbol{g}, \boldsymbol{\psi}^2)}[J_T^1(\Pi_T, U_T^2)],\end{aligned} \tag{44}$$

where, in the second equality, we expand the expectation in the LHS, substitute the definitions of $g_T^{*1}$ and $J_T^1$ from (39), and note that this yields the expectation in the RHS. □

Next, we show that, given the control law $g_T^{*1}$ for agent 1, the control law $g_T^{*2}$ is optimal for agent 2 at time $T$.

**Lemma 11.** *1) The value function $J_T^2$ in* (40) *is such that*

$$\mathcal{J}_T(\boldsymbol{g}) \geq \mathbb{E}^{(\boldsymbol{g}, \boldsymbol{\psi}^2)}[J_T^2(\Pi_T^2)], \quad \forall \boldsymbol{g} \in \mathcal{G}''. \tag{45}$$

*2) The corresponding control law $g_T^{*2}$ is such that*

$$\mathcal{J}_T(\boldsymbol{g}_{0:T-1}, g_T^{*1}, g_T^{*2}) = \mathbb{E}^{(\boldsymbol{g}, \boldsymbol{\psi}^2)}[J_T^2(\Pi_T^2)]. \tag{46}$$

*Proof.* 1) Agent 2 at time $T$ generates her action $U_T^2$ as a function of $\Pi_T^2$. Using Lemma 10, for all $\boldsymbol{g} \in \mathcal{G}''$,

$$\begin{aligned}\mathcal{J}_T(\boldsymbol{g}) &\geq \mathbb{E}^{(\boldsymbol{g}, \boldsymbol{\psi}^2)}[J_T^1(\Pi_T^1, U_T^2)] \\ &= \mathbb{E}^{(\boldsymbol{g}, \boldsymbol{\psi}^2)}\big[\mathbb{E}^{(\boldsymbol{g}, \boldsymbol{\psi}^2)}[J_T^1(\Pi_T^1, U_T^2) | \Pi_T^2, U_T^2]\big] \\ &\geq \mathbb{E}^{(\boldsymbol{g}, \boldsymbol{\psi}^2)}\big[J_T^2(\Pi_T^2)\big],\end{aligned} \tag{47}$$

where, in the equality, we use the law of iterated expectations, and in the second inequality, we use the definition of $J_T^2$ from (40).

2) Starting with the equality in (47), we substitute $U_T^1 = g_T^{*2}(\Pi_T^1)$ to write that

$$\begin{aligned}&\mathcal{J}_T(\boldsymbol{g}_{0:T-1}, g_T^{*1}, g_T^{*2}) \\ &= \mathbb{E}^{(\boldsymbol{g}, \boldsymbol{\psi}^2)}\big[\mathbb{E}^{(\boldsymbol{g}, \boldsymbol{\psi}^2)}[J_T^1(\Pi_T^1, g_T^{*2}(\Pi_T^2) \mid \Pi_T^2]\big] \\ &= \mathbb{E}^{(\boldsymbol{g}, \boldsymbol{\psi}^2)}\big[J_T^2(\Pi_T^2)\big],\end{aligned} \tag{48}$$

where, in the second equality, we expand the expectation in the LHS, substitute the definitions of $g_T^{*2}$ and $J_T^2$ from (40), and note that this yields the expectation in the RHS. □

Next, we show that the laws $(\psi_t^{*[2,1]}, g_t^{*2})$ are optimal for each $t = 0, \dots, T-1$.

**Theorem 3.** *For any $t = 0, \dots, T-1$:*

*1) The value function $J_t$ in (41) is such that*

$$\mathcal{J}_t(\boldsymbol{g}) \geq \mathbb{E}^{(\boldsymbol{g},\boldsymbol{\psi}^2)}\big[J_t(\Pi_t^2)\big], \quad \forall \boldsymbol{g} \in \mathcal{G}''. \tag{49}$$

*2) The corresponding laws $(\psi_t^{*[2,1]}, g_t^{*2})$ are such that*

$$\mathcal{J}_t(\boldsymbol{g}_{0:t-1}, g_t^{*1}, g_t^{*2}, , \boldsymbol{g}_{t+1:T}^*) = \mathbb{E}^{(\boldsymbol{g},\boldsymbol{\psi}^2)}[J_t(\Pi_t^2)], \tag{50}$$

*where $g_t^{*1}$ is derived from $\psi_t^{*[2,1]}$ using Lemma 3.*

*Proof.* Note the DP for time steps $t = 0, \dots, T-1$ is the same as a centralized DP for Problem 3. The optimality of such a DP can be proven using mathematical induction starting with time $T-1$ in a manner similar to [2], [18]. □

**Remark 2.** At time $T$, our DP has two sub-steps, each with a different value function. These sub-steps take advantage of the nested information structure to directly compute the control laws at time $T$, which involves solving two optimization problems with respect to control actions. This is simpler than solving an optimization problem involving a prescription of agent 2 for agent 1, as in time steps $t = 0, \dots, T-1$. Thus, our DP presents a simpler solution for the final time step.

## V. Discussion and Conclusions

In this paper, we introduced a dynamic team of two agents with a nested information structure and derived structural results for the optimal control strategies. Our derivation utilized a combination of the person-by-person and prescription approaches to arrive at a distinct structural form that cannot be achieved by either of the techniques alone. We also presented a DP that can be used to derive the optimal control strategies for a finite time horizon. Our DP utilized the nested information structure to simplify the computation of optimal control laws for the team at the final time step. Note that our results can be extended to teams of $n \in \mathbb{N}$ agents with nested information, by iteratively applying the person-by-person and the prescription approach. While our results do not yield a time invariant domain for optimal control strategies, their advantage is that they only require tracking probability distributions over finite valued supports. Thus, in comparison to related results in [15], [26], it may be easier to derive approximate strategies using our results. Furthermore, there may be systems with specific dynamics where we can divide the DP into multiple sub-steps at each time step. Our ongoing work seeks to extend our results to more general information structures and to investigate decentralized minimax control problems using these techniques like those reported in [27].

## References


[1] R. Radner, "Team decision problems," *The Annals of Mathematical Statistics*, vol. 33, no. 3, pp. 857–881, 1962.

[2] 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.

[3] 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).

[4] L. E. Beaver and A. A. Malikopoulos, "An Overview on Optimal Flocking," *Annual Reviews in Control*, vol. 51, pp. 88–99, 2021.

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

[6] Y.-C. Ho and K.-C. Chu, "Team decision theory and information structures in optimal control problems–Part I," *IEEE Transactions on Automatic Control*, vol. 17, no. 1, pp. 15–22, 1972.

[7] 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.

[8] A. Nayyar and L. Lessard, "Structural results for partially nested lqg systems over graphs," in *2015 American Control Conference (ACC)*, pp. 5457–5464, IEEE, 2015.

[9] S. Yuksel, "Stochastic nestedness and the belief sharing information pattern," *IEEE Transactions on Automatic Control*, vol. 54, no. 12, pp. 2773–2786, 2009.

[10] A. Dave and A. A. Malikopoulos, "Decentralized stochastic control in partially nested information structures," in *8th IFAC Workshop on Distributed Estimation and Control in Networked Systems*, 2019.

[11] 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.

[12] A. Nayyar, T. Başar, D. Teneketzis, and V. V. Veeravalli, "Optimal Strategies for Communication and Remote Estimation With an Energy Harvesting Sensor," *IEEE Transactions on Automatic Control*, vol. 58, no. 9, pp. 2246–2260, 2013.

[13] 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.

[14] A. A. Malikopoulos, "On team decision problems with nonclassical information structures," *arXiv:2101.10992*, 2021 (in review).

[15] 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.

[16] A. M. I. Mahbub and A. A. Malikopoulos, "A Platoon Formation Framework in a Mixed Traffic Environment," *IEEE Control Systems Letters (LCSS)*, vol. 6, pp. 1370–1375, 2021.

[17] A. Mahajan and S. Tatikonda, "An algorithmic approach to identify irrelevant information in sequential teams," *Automatica*, vol. 61, pp. 178–191, 2015.

[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] S. Yuksel, "A universal dynamic program and refined existence results for decentralized stochastic control," *SIAM Journal on Control and Optimization*, vol. 58, no. 5, pp. 2711–2739, 2020.

[20] 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.

[21] C. D. Charalambous, "Decentralized optimality conditions of stochastic differential decision problems via girsanov's measure transformation," *Mathematics of Control, Signals, and Systems*, vol. 28, no. 3, pp. 1–55, 2016.

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

[23] J. Subramanian and A. Mahajan, "Approximate information state for partially observed systems," in *2019 IEEE 58th Conference on Decision and Control (CDC)*, pp. 1629–1636, IEEE, 2019.

[24] H. Witsenhausen, "Some remarks on the concept of state," in *Directions in Large-Scale Systems*, pp. 69–75, Springer, 1976.

[25] A. Dave and A. A. Malikopoulos, "A dynamic program for a team of two agents with nested information," *preprint, arXiv:2103.10028*, 2021.

[26] J. Arabneydi and A. Mahajan, "Team optimal control of coupled subsystems with mean-field sharing," in *53rd IEEE Conference on Decision and Control*, pp. 1669–1674, Dec 2014.

[27] M. Gagrani and A. Nayyar, "Decentralized minimax control problems with partial history sharing," in *2017 American Control Conference (ACC)*, pp. 3373–3379, IEEE, 2017.