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

Common Information Belief based Dynamic Programs for Stochastic Zero-sum Games with Competing Teams

Dhruva Kartik, Ashutosh Nayyar and Urbashi Mitra

Abstract— **Decentralized team problems where players have asymmetric information about the state of the underlying stochastic system have been actively studied, but *games* between such teams are less understood. We consider a general model of zero-sum stochastic games between two competing teams. This model subsumes many previously considered team and zero-sum game models. For this general model, we provide bounds on the upper (min-max) and lower (max-min) values of the game. Furthermore, if the upper and lower values of the game are identical (i.e., if the game has a *value*), our bounds coincide with the value of the game. Our bounds are obtained using two dynamic programs based on a sufficient statistic known as the common information belief (CIB). We also identify certain information structures in which only the minimizing team controls the evolution of the CIB. In these cases, we show that one of our CIB based dynamic programs can be used to find the min-max strategy (in addition to the min-max value). We propose an approximate dynamic programming approach for computing the values (and the strategy when applicable) and illustrate our results with the help of an example.**

I. Introduction

In decentralized team problems, players collaboratively control a stochastic system to minimize a common cost. The information used by these players to select their control actions may be different. For instance, some of the players may have more information about the system state than others [1]; or each player may have some private observations that are shared with other players with some delay [2]. Such multi-agent team problems with an information asymmetry arise in a multitude of domains like autonomous vehicles and drones, power grids, transportation networks, military and rescue operations, wildlife conservation [3] etc. Over the past few years, several methods have been developed to address decentralized team problems [1], [4]–[7]. However, *games* between such teams are less understood. Many of the aforementioned systems are susceptible to adversarial attacks. Therefore, the strategies used by the team of players for controlling these systems must be designed in such a way that the damage inflicted by the adversary is minimized. Such adversarial interactions can be modeled as zero-sum games between competing teams, and our main goal in this paper is develop a framework that can be used to analyze and solve them.

Dhruva Kartik, Ashutosh Nayyar and Urbashi Mitra are with the Department of Electrical and Computer Engineering, University of Southern California, Los Angeles, CA, USA. Email: {mokhasun,ashutosn,ubli}@usc.edu. This research was supported by Grant ONR N00014-15-1-2550, NSF CCF-1817200, NSF ECCS 1750041, NSF CCF-2008927, ARO W911NF1910269, Cisco Foundation 1980393, ONR 503400-78050, DOE DE-SC0021417, Swedish Research Council 2018-04359 and Okawa Foundation.

The aforementioned works [1], [4]–[7] on cooperative team problems solve them by first constructing an auxiliary single-agent Markov Decision Process (MDP). The auxiliary state (state of the auxiliary MDP) is the *common information belief* (CIB). CIB is the belief on the system state and all the players' *private* information conditioned on the *common* (or public) information. Auxiliary actions (actions in the auxiliary MDP) are *mappings* from agents' private information to their actions [4]. The optimal values of the team problem and this auxiliary MDP are identical. Further, an optimal strategy for the team problem can be obtained using any optimal solution of the auxiliary MDP with a simple transformation. The optimal value and strategies of this auxiliary MDP (and thus the team problem) can be characterized by dynamic programs (a.k.a. Bellman equations or recursive formulas). A key consequence of this characterization is that the CIB is a sufficient statistic for optimal control in team problems. We investigate whether a similar approach can be used to characterize values and strategies in zero-sum games between teams. This extension is not straightforward. In general games (i.e., not necessarily zero-sum), it may not be possible to obtain such dynamic programs (DPs) and/or sufficient statistics [8], [9]. However, we show that for *zero-sum* games between teams, the values can be characterized by CIB based DPs. Further, we show that for some specialized models, the CIB based DPs can be used to characterize a min-max strategy as well. A key implication of our result is that this CIB based approach can be used to solve several team problems considered before [1], [4], [7] even in the presence of certain types of adversaries.

A phenomenon of particular interest and importance in team problems is *signaling*. Players in a team can agree upon their control strategies *ex ante*. Based on these agreed upon strategies, a player can often make inferences about the system state or the other players' private information (which are otherwise inaccessible to the player). This implicit form of communication between players is referred to as signaling and can be vital for effective coordination. While signaling is beneficial in cooperative teams, it can be detrimental in the presence of an adversary. This is because the adversary can exploit it to infer sensitive private information and inflict severe damage upon the system. A concrete example that illustrates this trade-off between *signaling* and *secrecy* is discussed is Section V. Our framework can be used optimize this trade-off in several stochastic games between teams.

a) Related Work on Games: Zero-sum games between two individual players with asymmetric information have been extensively studied. In [10]–[15], stochastic zero-sum

978-1-6654-5196-3/$31.00 ©2022 AACC

games with varying degrees of generality were considered and dynamic programming characterizations of the value of the game were provided. Various properties of the value functions (such as continuity) were also established and for some specialized information structures, these works also characterize a min-max strategy. Linear programs for computing the values and strategies in certain games were proposed in [16], [17]; and methods based on heuristic search value iteration (HSVI) [18] to compute the value of some games were proposed in [19], [20]. Zero-sum *extensive form* games in which a team of players competes against an adversary have been studied in [21]–[23]. Structured Nash equilibria in general games (i.e. not necessarily zero-sum) were studied in [24]–[26] under some assumptions on the system dynamics and players' information structure. A combination of reinforcement learning and search was used in [27] to solve two-player zero-sum games. While this approach has very strong empirical performance, a better analytical understanding of it is needed. Our work is closely related to [15], [28] and builds on their results. Our novel contributions in this paper over past works are summarized below.

b) Contributions: (i) In this paper, we study a *general* class of stochastic zero-sum games between two competing teams of players. Such team vs. team games present novel features because of the need for coordination and signaling within a team while preserving secrecy and minimizing losses against the opposing team. Our general model captures a variety of team vs team interactions that have not been studied before. In addition, our model covers previously considered settings in stochastic teams [1], [5], [7] and zero-sum games [15], [19], [20]. (ii) For our general model, we adapt the techniques in [15] to provide bounds on the upper (min-max) and lower (max-min) values of the game and characterize the value of the game when it exists. These bounds provide us with fundamental limits on the performance achievable by either team. Our bounds are obtained using two dynamic programs (DPs) based on a sufficient statistic known as the common information belief (CIB). (iii) We also identify a subclass of game models in which only one of the teams (say the minimizing team) controls the evolution of the CIB. In these cases, we show that one of our CIB based dynamic programs can be used to find the min-max value as well as a min-max *strategy*[1]. (iv) Our result reveals that the structure of the CIB based min-max strategy is similar to the structure of team optimal strategies. Such structural results have been successfully used in prior works [7], [27] to design efficient strategies for significantly challenging team problems. (v) Lastly, we discuss an approximate dynamic programming approach along with key structural properties for computing the values (and the strategy when applicable) and illustrate our results with the help of an example.

c) Notation: Random variables are denoted by upper case letters, their realizations by the corresponding lower case letters. In general, subscripts are used as time index while superscripts are used to index decision-making agents. For time indices $t_1 \leq t_2$, $X_{t_1:t_2}$ is the short hand notation for the variables $(X_{t_1}, X_{t_1+1}, ..., X_{t_2})$. Similarly, $X^{1:2}$ is the short hand notation for the collection of variables (X^1, X^2). Operators $\mathbb{P}(\cdot)$ and $\mathbb{E}[\cdot]$ denote the probability of an event, and the expectation of a random variable respectively. For random variables/vectors X and Y, $\mathbb{P}(\cdot|Y=y)$, $\mathbb{E}[X|Y=y]$ and $\mathbb{P}(X=x \mid Y=y)$ are denoted by $\mathbb{P}(\cdot|y)$, $\mathbb{E}[X|y]$ and $\mathbb{P}(x \mid y)$, respectively. For a strategy g, we use $\mathbb{P}^g(\cdot)$ (resp. $\mathbb{E}^g[\cdot]$) to indicate that the probability (resp. expectation) depends on the choice of g. For any finite set $\mathcal{A}$, $\Delta\mathcal{A}$ denotes the probability simplex over the set $\mathcal{A}$. For any two sets $\mathcal{A}$ and $\mathcal{B}$, $\mathcal{F}(\mathcal{A},\mathcal{B})$ denotes the set of all functions from $\mathcal{A}$ to $\mathcal{B}$. We define RAND to be mechanism that given (i) a finite set $\mathcal{A}$, (ii) a distribution d over $\mathcal{A}$ and a random variable K uniformly distributed over the interval $(0,1]$, produces a random variable $X \in \mathcal{A}$ with distribution d, i.e.,

$$X = \text{RAND}(\mathcal{A}, d, K) \sim d. \tag{1}$$

The rest of the paper is organized as follows. We formulate the problem in Section II. In Section III, we construct a virtual game using which bounds on the upper and lower values of the original game are characterized. In Section IV, we consider specialized models in which only the minimizing team controls the common information belief. For these models, we provide a tighter characterization of the upper value and also, a min-max strategy. In Section V, we discuss a computational approach for solving the dynamic program and illustrate it with the help of an example. The proofs of our results are provided in appendices, all of which are in [29].

II. Problem Formulation

Consider a dynamic system with two teams. Team 1 has N_1 players and Team 2 has N_2 players. The system operates in discrete time over a horizon[2] T. Let $X_t \in \mathcal{X}_t$ be the state of the system at time t, and let $U_t^{i,j} \in \mathcal{U}_t^{i,j}$ be the action of Player j, $j \in \{1,\dots,N_i\}$, in Team i, $i \in \{1,2\}$, at time t. Let

$$U_t^1 \doteq \left(U_t^{1,1},\dots,U_t^{1,N_1}\right); \quad U_t^2 \doteq \left(U_t^{2,1},\dots,U_t^{2,N_2}\right),$$

and $\mathcal{U}_t^i$ be the set of all possible realizations of U_t^i. We will refer to U_t^i as Team i's action at time t. The state of the system evolves in a controlled Markovian manner as

$$X_{t+1} = f_t(X_t, U_t^1, U_t^2, W_t^s), \tag{2}$$

where W_t^s is the system noise. There is an observation process $Y_t^{i,j} \in \mathcal{Y}_t^{i,j}$ associated with each Player j in Team i and is given as

$$Y_t^{i,j} = h_t^{i,j}(X_t, U_{t-1}^1, U_{t-1}^2, W_t^{i,j}), \tag{3}$$

where $W_t^{i,j}$ is the observation noise. Let us define

$$Y_t^1 \doteq \left(Y_t^{1,1},\dots,Y_t^{1,N_1}\right); \quad Y_t^2 \doteq \left(Y_t^{2,1},\dots,Y_t^{2,N_2}\right).$$

[1]Note that this characterization of a min-max strategy is *not* present in [15]. A similar result for a *very specific model with limited applicability* exists in [28]. Our result is substantially more general than that in [28].

[2]With a sufficiently large planning horizon, infinite horizon problems with discounted cost can be solved approximately as finite-horizon problems.

We assume that the sets $\mathcal{X}_t, \mathcal{U}_t^{i,j}$ and $\mathcal{Y}_t^{i,j}$ are finite for all i, j and t. Further, the random variables $X_1, W_t^s, W_t^{i,j}$ (referred to as *the primitive random variables*) can take finitely many values and are mutually independent.

Remark 1. An alternative approach commonly used for characterizing system dynamics and observation models is to specify the transition and observation probabilities. We emphasize that this alternative characterization is equivalent to ours in equations (2) and (3) [30].

a) Information Structure: At time t, Player j in Team i has access to *a subset of all observations and actions generated so far*. Let $I_t^{i,j}$ denote the collection of variables (i.e. observations and actions) available to Player j in team i at time t. Then $I_t^{i,j} \subseteq \cup_{i,j}\{Y_{1:t}^{i,j}, U_{1:t-1}^{i,j}\}$. The set of all possible realizations of $I_t^{i,j}$ is denoted by $\mathcal{I}_t^{i,j}$. Examples of such information structures include $I_t^{i,j} = \{Y_{1:t}^{i,j}, U_{1:t-1}^{i,j}\}$ which corresponds to the information structure in Dec-POMDPs [5] and $I_t^{i,j} = \{Y_{1:t}^{i,j}, Y_{1:t-d}^{1:2}, U_{1:t-1}^{1:2}\}$ wherein each player's actions are seen by all the players and their observations become public after a delay of d time steps.

Information $I_t^{i,j}$ can be decomposed into *common* and *private* information, i.e. $I_t^{i,j} = C_t \cup P_t^{i,j}$; common information C_t is the set of variables known to *all* players at time t. The private information $P_t^{i,j}$ for Player j in Team i is defined as $I_t^{i,j} \setminus C_t$. Let

$$P_t^1 \doteq \left(P_t^{1,1}, \ldots, P_t^{1,N_1}\right); \quad P_t^2 \doteq \left(P_t^{2,1}, \ldots, P_t^{2,N_2}\right).$$

We will refer to P_t^i as Team i's private information. Let $\mathcal{C}_t$ be the set of all possible realizations of common information at time t, $\mathcal{P}_t^{i,j}$ be the set of all possible realizations of private information for Player j in Team i at time t and $\mathcal{P}_t^i$ be the set of all possible realizations of P_t^i. We make the following assumption on the evolution of common and private information. This is similar to Assumption 1 of [15], [24].

Assumption 1. *The evolution of common and private information available to the players is as follows: (i) The common information C_t is non-decreasing with time, i.e. $C_t \subseteq C_{t+1}$. Let $Z_{t+1} \doteq C_{t+1} \setminus C_t$ be the increment in common information. Thus, $C_{t+1} = \{C_t, Z_{t+1}\}$. Furthermore,*

$$Z_{t+1} = \zeta_{t+1}(P_t^{1:2}, U_t^{1:2}, Y_{t+1}^{1:2}), \tag{4}$$

where ζ_{t+1} is a fixed transformation. (ii) The private information evolves as

$$P_{t+1}^i = \xi_{t+1}^i(P_t^{1:2}, U_t^{1:2}, Y_{t+1}^{1:2}), \tag{5}$$

where ξ_{t+1}^i is a fixed transformation and $i = 1, 2$.

As noted in [4], [15], a number of information structures satisfy the above assumption. Our analysis applies to any information structure that satisfies Assumption 1 including, among others, Dec-POMDPs and the delayed sharing information structure discussed above.

b) Strategies and Values: Players can use any information available to them to select their actions and we allow behavioral strategies for all players. Thus, at time t, Player j in Team i chooses a distribution $\delta U_t^{i,j}$ over its action space using a *control law* $g_t^{i,j} : \mathcal{I}_t^{i,j} \to \Delta\mathcal{U}_t^{i,j}$, i.e., $\delta U_t^{i,j} = g_t^{i,j}(I_t^{i,j}) = g_t^{i,j}(C_t, P_t^{i,j})$. The distrubtion $\delta U_t^{i,j}$ is then used to randomly generate the control action $U_t^{i,j}$ as follows. We assume that player j of Team i has access to i.i.d. random variables $K_{1:T}^{i,j}$ that are uniformly distributed over the interval $(0, 1]$. These uniformly distributed variables are independent of each other and of the primitive random variables. The action $U_t^{i,j}$ is generated using $K_t^{i,j}$ and the randomization mechanism described in (1), i.e.,

$$U_t^{i,j} = \text{RAND}(\mathcal{U}_t^{i,j}, \delta U_t^{i,j}, K_t^{i,j}). \tag{6}$$

The collection of control laws used by the players in Team i at time t is denoted by $g_t^i \doteq (g_t^{i,1}, \ldots, g_t^{i,N_i})$ and is referred to as the control law of Team i at time t. Let the set of all possible control laws for Team i at time t be denoted by $\mathcal{G}_t^i$. The collection of control laws $g^i \doteq (g_1^i, \ldots, g_T^i)$ is referred to as the *control strategy* of Team i, and the pair of control strategies (g^1, g^2) is referred to as a *strategy profile*. Let the set of all possible control strategies for Team i be $\mathcal{G}^i$.

The total expected cost associated with a strategy profile (g^1, g^2) is

$$J(g^1, g^2) \doteq \mathbb{E}^{(g^1, g^2)}\left[\sum_{t=1}^{T} c_t(X_t, U_t^1, U_t^2)\right], \tag{7}$$

where $c_t : \mathcal{X}_t \times \mathcal{U}_t^1 \times \mathcal{U}_t^2 \to \mathbb{R}$ is the cost function at time t. Team 1 wants to minimize the total expected cost, while Team 2 wants to maximize it. We refer to this zero-sum game between Team 1 and Team 2 as Game $\mathscr{G}$.

Definition 1. *The upper and lower values of the game $\mathscr{G}$ are respectively defined as*

$$S^u(\mathscr{G}) \doteq \min_{g^1 \in \mathcal{G}^1} \max_{g^2 \in \mathcal{G}^2} J(g^1, g^2), \tag{8}$$

$$S^l(\mathscr{G}) \doteq \max_{g^2 \in \mathcal{G}^2} \min_{g^1 \in \mathcal{G}^1} J(g^1, g^2). \tag{9}$$

If the upper and lower values are the same, they are referred to as the value of the game and denoted by $S(\mathscr{G})$. The minimizing strategy in (8) *is referred to as Team 1's optimal strategy and the maximizing strategy in* (9) *is referred to as Team 2's optimal strategy*[3].

A key objective of this work is to characterize the upper and lower values $S^u(\mathscr{G})$ and $S^l(\mathscr{G})$ of Game $\mathscr{G}$. To this end, we will define an *expanded* virtual game $\mathscr{G}_e$. This virtual game will be used to obtain bounds on the upper and lower values of the original game $\mathscr{G}$. These bounds happen to be tight when the upper and lower values of game $\mathscr{G}$ are equal. For a sub-class of information structures, we will show that the expanded virtual game $\mathscr{G}_e$ can be used to obtain optimal strategies for one of the teams.

[3]The strategy spaces $\mathcal{G}^1$ and $\mathcal{G}^2$ are compact and the cost $J(\cdot)$ is continuous in g^1, g^2. Hence, the existence of optimal strategies can be established using Berge's maximum theorem [31].

Remark 2. An alternative way of randomization is to use *mixed strategies* wherein a player randomly chooses a deterministic strategy at the beginning of the game and uses it for selecting its actions. According to Kuhn's theorem, mixed and behavioral strategies are equivalent when players have perfect recall [32].

Remark 3 (Independent and Shared Randomness). In most situations, the source of randomization is either privately known to the player (as in (6)) or publicly known to all the players in both teams. In this paper, we focus on independent randomization as in (6). In some situations, a shared source of randomness may be available to all players in Team i but not to any any of the players in the opposing team. Such shared randomness can help players in a team coordinate better. We believe that our approach can be extended to this case as well with some modifications.

We note that if the upper and lower values of game $\mathscr{G}$ are the same, then any pair of optimal strategies (g^{1*}, g^{2*}) forms a *Team Nash Equilibrium*[4], i.e., for every $g^1 \in \mathcal{G}^1$ and $g^2 \in \mathcal{G}^2$,

$$J(g^{1*}, g^2) \leq J(g^{1*}, g^{2*}) \leq J(g^1, g^{2*}).$$

In this case, $J(g^{1*}, g^{2*})$ is the value of the game, i.e. $J(g^{1*}, g^{2*}) = S^l(\mathscr{G}) = S^u(\mathscr{G}) = S(\mathscr{G})$. Conversely, if a Team Nash Equilibrium exists, then the upper and lower values are the same [34].

III. Expanded Virtual Game $\mathscr{G}_e$

The expanded virtual game $\mathscr{G}_e$ is constructed using the methodology in [15]. This game involves the same underlying system model as in game $\mathscr{G}$. The key distinction between games $\mathscr{G}$ and $\mathscr{G}_e$ lies in the manner in which the actions used to control the system are chosen. In game $\mathscr{G}_e$, all the players in each team of game $\mathscr{G}$ are replaced by a virtual player. Thus, game $\mathscr{G}_e$ has two virtual players, one for each team, and they operate as follows.

a) Prescriptions: Consider virtual player i associated with Team i, $i = 1, 2$. At each time t and for each $j = 1, \dots, N_i$, virtual player i selects a function $\Gamma_t^{i,j}$ that maps private information $P_t^{i,j}$ to a distribution $\delta U_t^{i,j}$ over the space $\mathcal{U}_t^{i,j}$. Thus, $\delta U_t^{i,j} = \Gamma_t^{i,j}(P_t^{i,j})$. The set of all such mappings is denoted by $\mathcal{B}_t^{i,j} \doteq \mathcal{F}(\mathcal{P}_t^{i,j}, \Delta\mathcal{U}_t^{i,j})$. We refer to the tuple $\Gamma_t^i \doteq (\Gamma_t^{i,1}, \dots, \Gamma_t^{i,N_i})$ of such mappings as virtual player i's *prescription* at time t. The set of all possible prescriptions for virtual player i at time t is denoted by $\mathcal{B}_t^i \doteq \mathcal{B}_t^{i,1} \times \cdots \times \mathcal{B}_t^{i,N_i}$. Once virtual player i selects its prescription, the action $U_t^{i,j}$ is randomly generated according to the distribution $\Gamma_t^{i,j}(P_t^{i,j})$. More precisely,

$$U_t^{i,j} = \textsc{rand}(\mathcal{U}_t^{i,j}, \Gamma_t^{i,j}(P_t^{i,j}), K_t^{i,j}), \tag{10}$$

where the random variable $K_t^{i,j}$ and the mechanism RAND are the same as in equation (6).

[4]When players in a team randomize independently, Team Nash equilirbia may not exist in general [33].

b) Strategies: The virtual players in game $\mathscr{G}_e$ have access to the common information C_t and all the past prescriptions of both players, i.e., $\Gamma_{1:t-1}^{1:2}$. Virtual player i selects its prescription at time t using a control law $\tilde{\chi}_t^i$, i.e., $\Gamma_t^i = \tilde{\chi}_t^i(C_t, \Gamma_{1:t-1}^{1:2})$. Let $\tilde{\mathcal{H}}_t^i$ be the set of all such control laws at time t and $\tilde{\mathcal{H}}^i \doteq \tilde{\mathcal{H}}_1^i \times \cdots \times \tilde{\mathcal{H}}_T^i$ be the set of all control strategies for virtual player i. The total cost for a strategy profile $(\tilde{\chi}^1, \tilde{\chi}^2)$ is

$$\mathcal{J}(\tilde{\chi}^1, \tilde{\chi}^2) = \mathbb{E}^{(\tilde{\chi}^1, \tilde{\chi}^2)} \left[\sum_{t=1}^{T} c_t(X_t, U_t^1, U_t^2) \right]. \tag{11}$$

The upper and lower values in $\mathscr{G}_e$ are defined as

$$S^u(\mathscr{G}_e) \doteq \min_{\tilde{\chi}^1 \in \tilde{\mathcal{H}}^1} \max_{\tilde{\chi}^2 \in \tilde{\mathcal{H}}^2} \mathcal{J}(\tilde{\chi}^1, \tilde{\chi}^2)$$
$$S^l(\mathscr{G}_e) \doteq \max_{\tilde{\chi}^2 \in \tilde{\mathcal{H}}^2} \min_{\tilde{\chi}^1 \in \tilde{\mathcal{H}}^1} \mathcal{J}(\tilde{\chi}^1, \tilde{\chi}^2).$$

The following theorem establishes the relationship between the upper and lower values of the expanded game $\mathscr{G}_e$ and the original game $\mathscr{G}$. This result is analogous to Theorem 1 from [15].

Theorem 1 (Proof in App. I)**.** *The lower and upper values of the two games described above satisfy the following: $S^l(\mathscr{G}) \leq S^l(\mathscr{G}_e) \leq S^u(\mathscr{G}_e) \leq S^u(\mathscr{G})$. Further, all these inequalities become equalities when a Team Nash equilibrium exists in Game $\mathscr{G}$.*

A. The Dynamic Programming Characterization

We describe a methodology for finding the upper and lower values of the expanded game $\mathscr{G}_e$ in this subsection. The results (and their proofs) in this subsection are similar to those in Section 4.2 of [15]. However, the prescription spaces $\mathcal{B}_t^i$ in this paper are different (and more general) from those in [15], and thus our results in this paper are more general. Our dynamic program is based on a sufficient statistic for virtual players in game $\mathscr{G}_e$ called the common information belief (CIB).

Definition 2. *At time t, the common information belief (CIB), denoted by Π_t, is defined as the virtual players' belief on the state and private information based on their information in game $\mathscr{G}_e$. Thus, for each $x_t \in \mathcal{X}_t, p_t^1 \in \mathcal{P}_t^1$ and $p_t^2 \in \mathcal{P}_t^2$, we have*

$$\Pi_t(x_t, p_t^{1:2}) \doteq \mathbb{P}\left[X_t = x_t, P_t^{1:2} = p_t^{1:2} \mid C_t, \Gamma_{1:t-1}^{1:2}\right].$$

The belief Π_t takes values in the set $\mathcal{S}_t \doteq \Delta(\mathcal{X}_t \times \mathcal{P}_t^1 \times \mathcal{P}_t^2)$.

The following lemma describes an update rule that can be used to compute the CIB.

Lemma 1 (Proof in App. II)**.** *For any strategy profile $(\tilde{\chi}^1, \tilde{\chi}^2)$ in Game $\mathscr{G}_e$, the common information based belief Π_t evolves almost surely as*

$$\Pi_{t+1} = F_t(\Pi_t, \Gamma_t^{1:2}, Z_{t+1}), \tag{12}$$

where F_t is a fixed transformation that does not depend on the virtual players' strategies. Further, the total expected cost

can be expressed as

$$\mathcal{J}(\tilde{\chi}^1, \tilde{\chi}^2) = \mathbb{E}^{(\tilde{\chi}^1, \tilde{\chi}^2)} \left[\sum_{t=1}^{T} \tilde{c}_t(\Pi_t, \Gamma_t^1, \Gamma_t^2) \right], \tag{13}$$

where the function $\tilde{c}_t$ is as defined in equation (58) in Appendix II.

a) Values in Game $\mathscr{G}_e$: We now describe two dynamic programs, one for each virtual player in $\mathscr{G}_e$. The minimizing virtual player (virtual player 1) in game $\mathscr{G}_e$ solves the following dynamic program. Define $V^u_{T+1}(\pi_{T+1}) = 0$ for every π_{T+1}. In a backward inductive manner, at each time $t \leq T$ and for each possible common information belief π_t and prescriptions γ_t^1, γ_t^2, define the upper cost-to-go function w_t^u and the upper value function V_t^u as

$$\begin{aligned} &w_t^u(\pi_t, \gamma_t^1, \gamma_t^2) \qquad (14)\\ &\doteq \tilde{c}_t(\pi_t, \gamma_t^1, \gamma_t^2) + \mathbb{E}[V^u_{t+1}(F_t(\pi_t, \gamma_t^{1:2}, Z_{t+1})) \mid \pi_t, \gamma_t^{1:2}],\\ &V_t^u(\pi_t) \doteq \min_{\gamma_t^1} \max_{\gamma_t^2} w_t^u(\pi_t, \gamma_t^1, \gamma_t^2). \qquad (15) \end{aligned}$$

The maximizing virtual player (virtual player 2) solves an analogous max-min dynamic program with a lower cost-to-go function w_t^l and lower value function V_t^l (See App. III for details).

Lemma 2 (Proof in App. III)**.** *For each t, there exists a measurable mapping $\Xi_t^1 : \mathcal{S}_t \to \mathcal{B}_t^1$ such that $V_t^u(\pi_t) = \max_{\gamma_t^2} w_t^u(\pi_t, \Xi_t^1(\pi_t), \gamma_t^2)$. Similarly, there exists a measurable mapping $\Xi_t^2 : \mathcal{S}_t \to \mathcal{B}_t^2$ such that $V_t^l(\pi_t) = \min_{\gamma_t^1} w_t^l(\pi_t, \gamma_t^1, \Xi_t^2(\pi_t))$.*

Theorem 2 (Proof in App. IV)**.** *The upper and lower values of the expanded virtual game $\mathscr{G}_e$ are given by $S^u(\mathscr{G}_e) = \mathbb{E}[V_1^u(\Pi_1)]$ and $S^l(\mathscr{G}_e) = \mathbb{E}[V_1^l(\Pi_1)]$.*

Theorem 2 gives us a dynamic programming characterization of the upper and lower values of the expanded game. As mentioned in Theorem 1, the upper and lower values of the expanded game provide bounds on the corresponding values of the original game. If the original game has a Team Nash equilibrium, then the dynamic programs described above characterize the value of the game.

b) Optimal Strategies in Game $\mathscr{G}_e$: The mappings Ξ^1 and Ξ^2 obtained from the dynamic programs described above (see Lemma 2) can be used to construct optimal strategies for both virtual players in game $\mathscr{G}_e$ in the following manner.

Definition 3. *Define strategies $\tilde{\chi}^{1*}$ and $\tilde{\chi}^{2*}$ for virtual players 1 and 2 respectively as follows: for each instance of common information c_t and prescription history $\gamma_{1:t-1}^{1:2}$, let*

$$\tilde{\chi}_t^{1*}(c_t, \gamma_{1:t-1}^{1:2}) \doteq \Xi_t^1(\pi_t); \quad \tilde{\chi}_t^{2*}(c_t, \gamma_{1:t-1}^{1:2}) \doteq \Xi_2^1(\pi_t),$$

where Ξ_t^1 and Ξ_t^2 are the mappings defined in Lemma 2 and π_t (which is a function of $c_t, \gamma_{1:t-1}^{1:2}$) is obtained in a forward inductive manner using the update rule F_t defined in Lemma 1.

Theorem 3 (Proof in App. IV)**.** *The strategies $\tilde{\chi}^{1*}$ and $\tilde{\chi}^{2*}$ as defined in Definition 3 are, respectively, min-max and max-min strategies in the expanded virtual game $\mathscr{G}_e$.*

IV. Only Virtual Player 1 controls the CIB

In this section, we consider a special class of instances of Game $\mathscr{G}$ and show that the dynamic program in (15) can be used to obtain a min-max *strategy* for Team 1, the minimizing team in game $\mathscr{G}$. The key property of the information structures considered in this section is that the common information belief Π_t is controlled[5] only by virtual player 1 in the corresponding expanded game $\mathscr{G}_e$. This is formally stated in the following assumption.

Assumption 2. *For any strategy profile $(\tilde{\chi}^1, \tilde{\chi}^2)$ in Game $\mathscr{G}_e$, the CIB Π_t evolves almost surely as*

$$\Pi_{t+1} = F_t(\Pi_t, \Gamma_t^1, Z_{t+1}), \tag{16}$$

where F_t is a fixed transformation that does not depend on the virtual players' strategies.

We will now describe some instances of Game $\mathscr{G}$ that satisfy Assumption 2. We note that two-player zero-sum games that satisfy a property similar to Assumption 2 were studied in [13].

A. Game Models Satisfying Assumption 2

a) All players in Team 2 have the same information: Consider an instance of game $\mathscr{G}$ in which every player j in Team 2 has the following information structure $I_t^{2,j} = \{Y_{1:t}^2, U_{1:t-1}^2\}$. Further, Team 2's information is known to every player in Team 1. Thus, the common information $C_t = I_t^{2,j}$. Under this condition, players in Team 2 do not have any private information. Thus, their private information $P_t^2 = \varnothing$. Any information structure satisfying the above conditions satisfies Assumption 2, see Appendix VI-A for a proof. Since Team 1's information structure is relatively unrestricted, the above model subsumes many previously considered team and game models. Notable examples of such models include: (i) all purely cooperative team problems in [1], [4], [7], [35], and (ii) two-player zero-sum game models where one agent is more informed than the other [13], [16], [19].

b) Team 2's observations become common information with one-step delay: Consider an an instance of game $\mathscr{G}$ where the current private information of Team 2 becomes common information in the very next time-step. More specifically, we have $C_{t+1} \supseteq \{Y_{1:t}^2, U_{1:t}^2\}$ and for each Player j in Team 2, $P_t^{2,j} = Y_t^{2,j}$. Note that unlike in [16], [19], players in Team 2 have some private information in this model. Any information structure that satisfies the above conditions satisfies Assumption 2, see Appendix VI-B for a proof.

[5]Note that the players in Team 2 might still be able to control the state dynamics through their actions.

Algorithm 1 Strategy $g^{1,j*}$ for Player j in Team 1

Input: $\Xi^1_t(\pi)$ obtained from DP for all t and all π
for $t = 1$ **to** T **do**
  Current information: $C_t, P^{1,j}_t$ {where $C_t = \{C_{t-1}, Z_t\}$}
  Update CIB $\Pi_t = F_{t-1}(\Pi_{t-1}, \Xi^1_{t-1}(\Pi_{t-1}), Z_t)$ {If $t = 1$, Initialize CIB Π_t using C_t}
  Get prescription $\Gamma^1_t = (\Gamma^{1,1}_t, \dots, \Gamma^{1,N_1}_t) = \Xi^1_t(\Pi_t)$
  Get distribution $\delta U^{1,j}_t = \Gamma^{1,j}_t(P^{1,j}_t)$ and select action $U^{1,j}_t = \text{RAND}(\mathcal{U}^{1,j}_t, \delta U^{1,j}_t, K^{1,j}_t)$
end for

c) Games with symmetric information: Consider the information structure where $I^{i,j}_t = \cup_{i,j}\{Y^{i,j}_{1:t}, U^{i,j}_{1:t-1}\}$ for every i, j. All the players in this game have the same information and thus, players do not have any private information. Note that this model subsumes perfect information games. It can be shown that this model satisfies Assumption 2 using the same arguments in Appendix VI-A. In this case, the CIB is not controlled by both virtual players and thus, we can use the dynamic program to obtain both min-max and max-min strategies.

In addition to the models discussed above, there are other instances of $\mathscr{G}$ that satisfy Assumption 2. These are included in Appendix V.

B. Min-max Value and Strategy in Game $\mathscr{G}$

a) Dynamic Program: Since we are considering special cases of Game $\mathscr{G}$, we can use the analysis in Section III to write the min-max dynamic program for virtual player 1. Because of Assumption 2, the belief update $F_t(\pi_t, \gamma^{1:2}_t, z_{t+1})$ in (14) is replaced by $F_t(\pi_t, \gamma^1_t, z_{t+1})$. Using Theorems 2 and 3, we can can conclude that the upper value of the expanded game $S^u(\mathscr{G}_e) = \mathbb{E}[V^u_1(\Pi_1)]$ and that the strategy $\tilde{\chi}^{1*}$ obtained from the DP is a min-max strategy for virtual player 1 in Game $\mathscr{G}_e$. An approximate dynamic programming based approach for solving the dynamic programs is discussed in Appendix VIII. This discussion includes certain structural properties of the value functions that make their computation significantly more tractable.

b) Min-max Value and Strategy: The following results provide a characterization of the min-max value $S^u(\mathscr{G})$ and a min-max strategy g^{1*} in game $\mathscr{G}$ under Assumption 2. Note that unlike the inequality in Theorem 1, the upper values of games $\mathscr{G}$ and $\mathscr{G}_e$ are always equal in this case.

Theorem 4 (Proof in App. VII)**.** *Under Assumption 2, we have* $S^u(\mathscr{G}) = S^u(\mathscr{G}_e) = \mathbb{E}[V^u_1(\Pi_1)]$.

Theorem 5 (Proof in App. VII)**.** *Under Assumption 2, the strategy* g^{1*} *defined in Algorithm 1 is a min-max strategy for Team 1 in the original game* $\mathscr{G}$.

V. A Special Case and Numerical Experiments

Consider an instance of Game $\mathscr{G}$ in which Team 1 has two players and Team 2 has only one player. At each time t, Player 1 in Team 1 observes the state perfectly, i.e. $Y^{1,1}_t = X_t$, but the player in Team 2 gets an imperfect observation Y^2_t defined as in (3). Player 1 has complete information: at each time t, it knows the entire state, observation and action histories of all the players. The player in Team 2 has partial information: at each time t, it knows its observation history $Y^2_{1:t}$ and action histories of all the players. Player 2 in Team 1 has the same information as that of the player in Team 2. Thus, the total information available to each player at t is as follows:

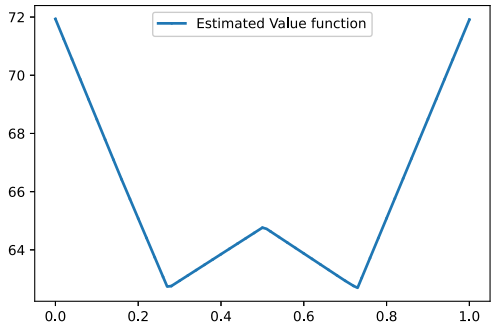
$$I^{1,1}_t = \{X_{1:t}, Y^2_{1:t}, U^{1:2}_{1:t-1}\}; \quad I^{1,2}_t = I^2_t = \{Y^2_{1:t}, U^{1:2}_{1:t-1}\}.$$

Clearly, $I^2_t \subseteq I^{1,1}_t$. The common and private information for this game can be written as follows: $C_t = I^2_t$, $P^{1,1}_t = \{X_{1:t}\}$ and $P^{1,2}_t = P^2_t = \varnothing$. The increment in common information at time t is $Z_t = \{Y^2_t, U^{1:2}_{t-1}\}$. In the game described above, the private information in $P^{1,1}_t$ includes the entire state history. However, Player 1 in Team 1 can ignore the past states $X_{1:t-1}$ without loss of optimality.

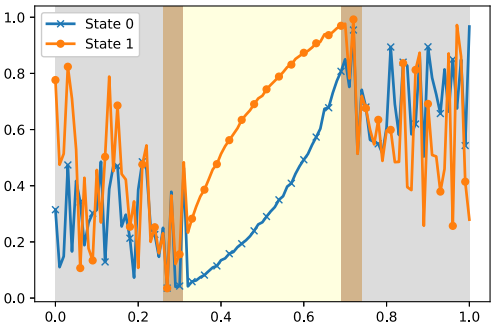
Lemma 3 (Proof in App. IX)**.** *There exists a min-max strategy* g^{1*} *such that the control law* $g^{1,1*}_t$ *at time* t *uses only* X_t *and* I^2_t *to select* $\delta U^{1,1}_t$*, i.e.,* $\delta U^{1,1}_t = g^{1,1*}_t(X_t, I^2_t)$.

The lemma above implies that, for the purpose of characterizing the value of the game and a min-max strategy for Team 1, we can restrict player 1's information structure to be $I^{1,1}_t = \{X_t, I^2_t\}$. Thus, the common and private information become: $C_t = I^2_t$, $P^{1,1}_t = \{X_t\}$ and $P^2_t = P^{1.1}_t = \varnothing$. We refer to this game with reduced private information as Game $\mathscr{H}$. The corresponding expanded virtual game is denoted by $\mathscr{H}_e$. A general methodology for reducing private information in decentralized team and game problems can be found in [36]. The information structure in $\mathscr{H}$ is a special case of the first information structure in Section IV-A, and thus satisfies Assumption 2. Therefore, using the dynamic program in Section IV-B, we can obtain the value function V^u_1 and the min-max strategy g^{1*}.

a) Numerical Experiments: We consider a particular example of game $\mathscr{H}$ described above. In this example, there are two entities (l and r) that can potentially be attacked and at any given time, exactly one of the entities is vulnerable. Player 1 of Team 1 knows which of the two entities is vulnerable whereas all the other players do not have this information. Player 2 of Team 1 can choose to *defend* one of the entities. The attacker in Team 2 can either launch a *blanket attack* on both entities or launch a *targeted attack* on one of the entities. When the attacker launches a blanket attack, the damage incurred by the system is minimal if Player 2 in Team 1 happens to be defending the vulnerable entity and the damage is substantial otherwise. When the attacker launches a targeted attack on the vulnerable entity, the damage is substantial irrespective of the defender's position. But if the attacker targets the invulnerable entity, the attacker becomes passive and cannot attack for some time. Thus, launching a targeted attack involves high risk for the attacker. The state of the attacker (active or passive) and all the players' actions are public information. The system state



(a) An estimate of the value function $V_1^u(\cdot)$.



(b) Prescriptions at $t = 1$ for Player 1 in Team 1.

Fig. 1. In these plots, the x-axis represents $\pi_1(0)$ and we restrict our attention to those beliefs where $\pi_1(0) + \pi_1(1) = 1$, i.e. when the attacker is active. In Figure 1(b), the blue and red curves respectively depict the Bernoulli probabilities associated with the distributions $\gamma_1^{1,1}(0)$ and $\gamma_1^{1,1}(1)$, where $\gamma_1^{1,1}$ is Player 1's prescription in Team 1.

X_t thus has two components, the hidden state (l or r) and the state of the attacker (a or p). For convenience, we will denote the states (l, a) and (r, a) with 0 and 1 respectively.

The only role of Player 1 in Team 1 in this game is to signal the hidden state using two available actions α and β. The main challenge is that both the defender and the attacker can see Player 1's actions. Player 1 needs to signal the hidden state to some extent so that its teammate's defense is effective under blanket attacks. However, if too much information is revealed, the attacker can exploit it to launch a targeted attack and cause significant damage. In this example, the key is to design a strategy that can balance between these two contrasting goals of *signaling* and *secrecy*. A precise description of this model is provided in Appendix X.

In order to solve this problem, we used the approximate DP approach discussed in Appendix VIII. The value function $V_1^u(\cdot)$ thus obtained is shown in Figure 1(a). The tension between signaling and secrecy can be seen in the shape of the value function in Figure 1(a). When the CIB $\pi_1(0) = 0.5$, the value function is concave in its neighborhood and decreases as we move away from 0.5. This indicates that in these belief states, revealing the hidden state to some extent is preferable. However, as the belief goes further away from 0.5, the value function starts increasing at some point. This indicates that the adversary has too much information and is using it to inflict damage upon the system. Figure 1(b) depicts Player 1's prescriptions leading to non-trivial signaling patterns at various belief states. Notice that the distributions $\gamma_1^{1,1}(0)$ and $\gamma_1^{1,1}(1)$ for hidden states 0 and 1 are quite distinct when $\pi_1(0) = 0.5$ (indicating significant signaling) and are nearly identical when $\pi_1(0) = 0.72$ (indicating negligible signaling). A more detailed discussion on our experimental results can be found in Appendix X.

VI. Conclusions

We considered a general model of stochastic zero-sum games between two competing decentralized teams and provided bounds on their upper and lower values in the form of CIB based dynamic programs. When game has a value, our bounds coincide with the value. We identified several instances of this game model (including previously considered models) in which the CIB is controlled only by one of the teams (say the minimizing team). For such games, we also provide a characterization of the min-max strategy. Under this strategy, each player only uses the current CIB and its private information to select its actions. The sufficiency of the CIB and private information for optimality can potentially be exploited to design efficient strategies in various problems. Finally, we proposed a computational approach for approximately solving the CIB based DPs. There is significant scope for improvement in our computational approach. Tailored forward exploration heuristics for sampling the belief space and adding a policy network can improve the accuracy and tractability of our approach.

References

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

[2] Ashutosh Nayyar, Aditya Mahajan, and Demosthenis Teneketzis, "Optimal control strategies in delayed sharing information structures," *IEEE Transactions on Automatic Control*, vol. 56, no. 7, pp. 1606–1620, 2010.

[3] Fei Fang, Milind Tambe, Bistra Dilkina, and Andrew J Plumptre, *Artificial intelligence and conservation*, Cambridge University Press, 2019.

[4] Ashutosh Nayyar, Aditya Mahajan, and Demosthenis 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.

[5] Frans A Oliehoek and Christopher Amato, *A concise introduction to decentralized POMDPs*, Springer, 2016.

[6] Tabish Rashid, Mikayel Samvelyan, Christian Schroeder, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson, "Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning," in *International Conference on Machine Learning*. PMLR, 2018, pp. 4295–4304.

[7] Jakob Foerster, Francis Song, Edward Hughes, Neil Burch, Iain Dunning, Shimon Whiteson, Matthew Botvinick, and Michael Bowling, "Bayesian action decoder for deep multi-agent reinforcement learning," in *International Conference on Machine Learning*. PMLR, 2019, pp. 1942–1951.

[8] Ashutosh Nayyar and Abhishek Gupta, "Information structures and values in zero-sum stochastic games," in *American Control Conference (ACC), 2017*. IEEE, 2017, pp. 3658–3663.

[9] Dengwang Tang, Hamidreza Tavafoghi, Vijay Subramanian, Ashutosh Nayyar, and Demosthenis Teneketzis, "Dynamic games among teams with delayed intra-team information sharing," *arXiv preprint arXiv:2102.11920*, 2021.

[10] Jean-François Mertens, Sylvain Sorin, and Shmuel Zamir, *Repeated games*, vol. 55, Cambridge University Press, 2015.

[11] Dinah Rosenberg, Eilon Solan, and Nicolas Vieille, "Stochastic games with a single controller and incomplete information," *SIAM journal on control and optimization*, vol. 43, no. 1, pp. 86–110, 2004.

[12] Jérôme Renault, "The value of repeated games with an informed controller," *Mathematics of operations Research*, vol. 37, no. 1, pp. 154–179, 2012.

[13] Fabien Gensbittel, Miquel Oliu-Barton, and Xavier Venel, "Existence of the uniform value in zero-sum repeated games with a more informed controller," *Journal of Dynamics and Games*, vol. 1, no. 3, pp. 411–445, 2014.

[14] Xiaoxi Li and Xavier Venel, "Recursive games: uniform value, tauberian theorem and the mertens conjecture," *International Journal of Game Theory*, vol. 45, no. 1-2, pp. 155–189, 2016.

[15] Dhruva Kartik and Ashutosh Nayyar, "Upper and lower values in zero-sum stochastic games with asymmetric information," *Dynamic Games and Applications*, pp. 1–26, 2020.

[16] Jiefu Zheng and David A Castañón, "Decomposition techniques for Markov zero-sum games with nested information," in *52nd IEEE Conference on Decision and Control*. IEEE, 2013, pp. 574–581.

[17] Lichun Li and Jeff Shamma, "LP formulation of asymmetric zero-sum stochastic games," in *53rd IEEE Conference on Decision and Control*. IEEE, 2014, pp. 1930–1935.

[18] Trey Smith and Reid Simmons, "Heuristic search value iteration for pomdps," in *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, 2004, pp. 520–527.

[19] Karel Horák, Branislav Bošanskỳ, and Michal Pěchouček, "Heuristic search value iteration for one-sided partially observable stochastic games," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017, vol. 31.

[20] Karel Horák and Branislav Bošanskỳ, "Solving partially observable stochastic games with public observations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, pp. 2029–2036.

[21] Bernhard von Stengel and Daphne Koller, "Team-maxmin equilibria," *Games and Economic Behavior*, vol. 21, no. 1-2, pp. 309–321, 1997.

[22] Gabriele Farina, Andrea Celli, Nicola Gatti, and Tuomas Sandholm, "Ex ante coordination and collusion in zero-sum multiplayer extensive-form games," in *Conference on Neural Information Processing Systems (NIPS)*, 2018.

[23] Youzhi Zhang and Bo An, "Computing team-maxmin equilibria in zero-sum multiplayer extensive-form games," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, pp. 2318–2325.

[24] Ashutosh Nayyar, Abhishek Gupta, Cedric Langbort, and Tamer Başar, "Common information based Markov perfect equilibria for stochastic games with asymmetric information: Finite games," *IEEE Transactions on Automatic Control*, vol. 59, no. 3, pp. 555–570, 2014.

[25] Yi Ouyang, Hamidreza Tavafoghi, and Demosthenis Teneketzis, "Dynamic games with asymmetric information: Common information based perfect bayesian equilibria and sequential decomposition," *IEEE Transactions on Automatic Control*, vol. 62, no. 1, pp. 222–237, 2017.

[26] Deepanshu Vasal, Abhinav Sinha, and Achilleas Anastasopoulos, "A systematic process for evaluating structured perfect bayesian equilibria in dynamic games with asymmetric information," *IEEE Transactions on Automatic Control*, vol. 64, no. 1, pp. 78–93, 2019.

[27] Noam Brown, Anton Bakhtin, Adam Lerer, and Qucheng Gong, "Combining deep reinforcement learning and search for imperfect-information games," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds. 2020, vol. 33, pp. 17057–17069, Curran Associates, Inc.

[28] Dhruva Kartik and Ashutosh Nayyar, "Stochastic zero-sum games with asymmetric information," in *58th IEEE Conference on Decision and Control*. IEEE, 2019.

[29] Dhruva Kartik, Ashutosh Nayyar, and Urbashi Mitra, "Common information belief based dynamic programs for stochastic zero-sum games with competing teams," *arXiv preprint arXiv:2102.05838*, 2021.

[30] Panganamala Ramana Kumar and Pravin Varaiya, *Stochastic systems: Estimation, identification, and adaptive control*, vol. 75, SIAM, 2015.

[31] A Hitchhiker's Guide, *Infinite dimensional analysis*, Springer, 2006.

[32] Michael Maschler, Eilon Solan, and Shmuel Zamir, *Game Theory*, Cambridge University Press, 2013.

[33] Venkat Anantharam and Vivek Borkar, "Common randomness and distributed control: A counterexample," *Systems & control letters*, vol. 56, no. 7-8, pp. 568–572, 2007.

[34] Martin J Osborne and Ariel Rubinstein, *A course in game theory*, MIT press, 1994.

[35] Jilles Steeve Dibangoye, Christopher Amato, Olivier Buffet, and François Charpillet, "Optimally solving dec-pomdps as continuous-state mdps," *Journal of Artificial Intelligence Research*, vol. 55, pp. 443–497, 2016.

[36] Hamidreza Tavafoghi, Yi Ouyang, and Demosthenis Teneketzis, "A sufficient information approach to decentralized decision making," in *2018 IEEE Conference on Decision and Control (CDC)*. IEEE, 2018, pp. 5069–5076.

[37] Onésimo Hernández-Lerma and Jean B Lasserre, *Discrete-time Markov control processes: basic optimality criteria*, vol. 30, Springer Science & Business Media, 2012.

[38] Dimitri P Bertsekas and John N Tsitsiklis, *Neuro-dynamic programming*, vol. 5, Athena Scientific Belmont, MA, 1996.

[39] Tianyi Lin, Chi Jin, and Michael Jordan, "On gradient descent ascent for nonconvex-concave minimax problems," in *International Conference on Machine Learning*. PMLR, 2020, pp. 6083–6093.

[40] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *arXiv preprint arXiv:1706.08500*, 2017.