2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

Coordination in Markov Games with Asymmetric Information

Xupeng Wei, Achilleas Anastasopoulos

Abstract— **We study coordination in Markov games with asymmetric information. We consider a model where the state consists of different components, each representing the private type of each player. Players' actions depend on their private types and the public observation of past actions. The state components evolve as independent Markov processes conditioned on actions. We propose a solution concept called perfect correlated equilibrium (PCE), realized by a correlation device that observes only the public information of past actions. At time t, the device generates a prescription profile from a commonly known joint distribution, and sends each player a prescription privately before they act. Players are expected to take actions according to the prescriptions at equilibrium by evaluating the suggested prescription at the private types. We introduce "structured" PCE (sPCE), in which the correlation device generates prescriptions based on the common action history through a common belief on the state. We motivate sPCE by showing that any payoff profile induced by a general device can be induced by a structured one. We show that when the correlation device is using structured strategies, players' rationality constraints can be characterized through appropriate Markov decision processes (MDPs). Based on this characterization, we develop a backward dynamic approach, with which one can verify if a structured device is feasible, or even design a structured PCE in a backward recursive manner. Finally, we consider a specific example demonstrating how coordination can improve social welfare.**

I. Introduction

Coordination in multi-agent systems with asymmetric information has been extensively studied in the literature on decentralized stochastic control [1]–[3] in the context of dynamic teams. However, when it comes to a system with strategic agents (as in a game setting), owing to the heterogeneous objectives and partial information, the classical approaches from non-strategic stochastic control cannot apply. Recent works [4], [5] studied variations of perfect Bayesian equilibrium (PBE) under the non-cooperative framework. Since PBE does not coordinate agents' actions, it can lead to a set of unfavorable equilibria resulting in lower social welfare than that of the team problem, and in some cases, even the non-existence of equilibrium [6]. An effective way to mitigate the loss of social welfare caused by strategic behaviour is introduction of some form of coordination.

Correlated equilibrium introduced in [7] is an appropriate solution concept to incorporate the coordination aspect in game settings. A correlated equilibrium is realized by a correlation device, which coordinates players by sending private but correlated signals (generated by a commonly known joint distribution) to players. The seminal work of [7] focuses on static games with complete information, and the idea is naturally extended to cases with asymmetric information in [8], [9], in which agents do not share a common observation and need to form a belief on unobservable parts in order to predict what others will do. Another line of works apply correlated equilibria to dynamic decision processes [10], especially to Markov games [11]. Some works combine dynamic aspects with information asymmetry. For example, [12], [13] look into correlated equilibrium in Markov games with asymmetric observations on states or actions, while they do not consider agents' rationality on information sets off the equilibrium paths (known as *subgame perfection*). Works such as [14], [15] investigate a general extensive form correlated equilibrium, but these general devices can suffer from a complexity that increases exponentially with the time horizon.

Despite a variety of forms of correlated equilibrium from previous works, we point out an inherent restriction on the communication ability of correlation devices in the case of Markov games with asymmetric information. Due to privacy concerns, agents in such an environment may not be willing to reveal their private types to the coordinator. To mitigate the privacy issue, in this paper, correlation devices are required to realize the coordination without utilizing private information. Essentially the correlation devices have to instruct users how to form an action based on their private information, instead of instructing them directly what action to take.

We study coordination in Markov games with independent asymmetric information. In our Markov game model, the state at each time consists of components that evolve independently conditioned on actions taken by all players. Each player privately observes one component of the state as her private type, and chooses an action based on private observations as well as the previous actions that are publicly observable. Instantaneous rewards of players are determined by the current state and actions. The main contributions of this paper are summarized as follows:

- We propose a new framework of coordination for Markov games with asymmetric information, in which the correlation device does not have access to agents' private types. Instead, it generates "prescriptions" from a commonly known joint distribution, and sends each player a prescription privately that instructs the agent what action to take given her private type. Under this framework, we develop a concept that we call *perfect correlated equilibrium* (PCE) that possesses the subgame perfection property.
- We introduce a "structured" PCE (sPCE) using a structured device that relies on common history through

This research is funded by the National Science Foundation grant ECCS 2015191.

a common belief on the current state. To motivate the structured device and equilibrium, we show that structured devices can achieve all the payoff profiles achieved by general devices, and the use of structured devices simplifies players' rationality constraints.
- We develop a backward dynamic programming approach for sPCE. By this approach, one may verify if a structured device induces an sPCE in a systematic and tractable manner, or even design an sPCE.

The remainder of the paper is organized as follows. Section II introduces the Markov game model with coordination. We propose PCE as a solution concept for coordination in Section III. Section IV motivates and investigates sPCE, as a subset of PCE, and then develops a backward recursive characterization for sPCE. To further illustrate the idea of sPCE, we study a concrete example in Section V. We conclude this paper in Section VI.

We use upper case letters to denote random variables, use the corresponding lower case letters to denote their realizations, and its calligraphy font as the set of possible realizations, except for T (time horizon) and N (number of players)s, Q as probability kernels, $\boldsymbol{\mathcal{G}}, \boldsymbol{\mathcal{P}}$ as the sets for variables $\boldsymbol{\gamma}, \boldsymbol{\pi}$. In notation x_t^i, the superscript indicates it is a variable for player i; the subscript indicates the variable is realized at time t. Define shorthands $x_{1:t} = (x_1, \dots, x_t)$, $\boldsymbol{x} = x^{1:N} = (x^1, \dots, x^N)$, and $x^{-i} = (x^1, \dots, x^{i-1}, x^{i+1}, \dots, x^N)$. $Q(\cdot)$ and $Q(\cdot|\cdot)$ represent pre-specified probability kernels; $\mathbb{P}_\mu^\beta$ and $\mathbb{E}_\mu^\beta$ represent probability and expectation under a given strategy profile β and belief system μ (to be defined later). We use the shorthand $\mathbb{P}_\mu^\beta(x|y) := \mathbb{P}_\mu^\beta(X = x|Y = y)$ when there is no ambiguity. The superscript β and subscript μ can be dropped if the quantity does not depend on them. $\Delta(\mathcal{X}) := \{\boldsymbol{p} \in [0,1]^{\mathcal{X}} | \sum_{x\in\mathcal{X}} p_x = 1\}$. For functions with the form $f : \mathcal{Y} \mapsto \Delta(\mathcal{X})$, use $f[y]$ to denote $f(\cdot|y)$. $1_a(x)$ is an indicator function, which returns 1 if $x = a$, and returns 0 otherwise.

II. Model

We consider a Markov game with N strategic players in $\mathcal{N} = \{1, \dots, N\}$ and a finite time horizon T. At each time t, the system state is $\boldsymbol{x}_t = x_t^{1:N}$, where $x_t^i \in \mathcal{X}^i$ is private information of player i. Player i at time t chooses an action $a_t^i \in \mathcal{A}^i$. It is assumed that both $\mathcal{X}^i$ and $\mathcal{A}^i$ are finite sets for all i. State components are independent conditioned on $a_{1:t}^{1:N}$ and their statistics are determined by the probability kernel

$$Q(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t, \boldsymbol{a}_t) = \prod_{i=1}^N Q^i(x_{t+1}^i|x_t^i, \boldsymbol{a}_t),\ t \geq 1, \quad (1)$$

$$Q(\boldsymbol{x}_1) = \prod_{i=1}^N Q^i(x_1^i). \quad (2)$$

Player i's instantaneous reward at time t is $r_t^i(\boldsymbol{x}_t, \boldsymbol{a}_t)$. A strategic player i would choose an appropriate strategy to maximize the expected value of the total reward from $t = 1$ to T.

A *coordinator* is a third-party that participates in the Markov game by committing to a *correlation device*. The correlation device influences players' behavior by sending them suggestions in the form of *prescriptions* $\gamma_t^i \in \mathcal{G}^i$:

$$\gamma_t^i : \mathcal{X}^i \mapsto \mathcal{A}^i. \quad (3)$$

A prescription γ_t^i is a suggestion for what player i should do given each possible observation x_t^i. At each time t, before players take action $\boldsymbol{a}_t$, the correlation device first broadcasts the past prescription profile $\boldsymbol{\gamma}_{t-1}$, and then generates a new prescription profile $\boldsymbol{\gamma}_t$ based on common observations. Subsequently it sends the new prescription γ_t^i to each player i privately. Therefore, at time t, the common observation $h_t^C \in \mathcal{H}_t^C$ between all players and the coordinator is

$$h_t^C := (\boldsymbol{a}_{1:t-1}, \boldsymbol{\gamma}_{1:t-1}), \quad (4)$$

and the private observation $h^{P,i} \in \mathcal{H}^{P,i}$ for player i is

$$h_t^{P,i} := (x_{1:t}^i, \gamma_t^i). \quad (5)$$

Denote the coordinator strategy as $\phi^C = \phi_{1:T}^C$, where

$$\phi_t^C : \mathcal{H}_t^C \mapsto \Delta(\boldsymbol{\mathcal{G}}), \quad (6)$$

which is common knowledge among all players due to the coordinator's commitment. Notice that ϕ_t^C's input is a common observation, and its output is a distribution over prescription profiles $\boldsymbol{\gamma}_t = \gamma_t^{1:N}$, instead of a product of distributions on single prescription γ_t^i. The fact that prescription suggestions $\gamma_t^1, \dots, \gamma_t^N$ are generated in dependent way is the means by which coordination among agents is achieved.

In this work, the coordinator's goal is to design a correlation device ϕ^C, such that a rational player will always follow the prescription provided by ϕ^C, i.e., each player i at time t will prefer to play $a_t^i = \gamma_t^i(x_t^i)$ over any other action $a_t^{'i} = g_t^i(x_{1:t}^i, \boldsymbol{a}_{1:t-1}, \boldsymbol{\gamma}_{1:t-1})$ (this behaviour is known as known as *obedience*).

Before wrapping up this section, we bring up two remarks regarding the coordinator. First, we restrict attention to deterministic prescriptions, as opposed to the more general case of randomized prescriptions $\gamma_t^i : \mathcal{X}^i \mapsto \Delta(\mathcal{A}^i)$ used in [5] for perfect Bayesian equilibria. Though this simplification restricts the candidate set of prescriptions to a tractable finite space, in later sections we will find that the set of reachable beliefs is also restricted to be finite, which leaves less freedom for the correlation device designer. We emphasize however that the results in this paper can be easily extended to the case with randomized prescriptions (which may come from a subset of all possible randomized prescriptions), with minor modifications in the derivations. Second, we emphasize that the proposed coordinator in this paper is not a centralized controller. The most essential difference that distinguishes a coordinator from a centralized controller is that the coordinator has no access to the complete information from agents. Hence, it is challenging for a coordinator to persuade strategic agents with private information.

III. Perfect Correlated Equilibria

Now that we have introduced the model of a Markov game with asymmetric information and a coordinator, we propose a solution concept that predicts the outcome of

the coordination, namely PCE. Essentially, to define such a solution concept, one needs to specify the details of the players' rationality, including how they form beliefs toward unknown variables, and what decision making problems they are actually facing. The PCE is then defined by the constraints implied by players' rationality.

A. Consistent Belief System

We investigate how the players form their beliefs toward unobservable payoff-relevant random variables, given a fixed correlation device ϕ^C. At time t, the instantaneous reward or future reward of player i depends on the unobserved x^{-i}. To compute this quantity, player i needs to form a belief on x_t^{-i}. For this purpose, we introduce a belief system μ which takes the observation $(h_t^C, h_t^{P,i})$ of user i at time t, and produces a distribution on $\mathcal{X}^{-i}$, i.e., the private belief of user i at time t has the form

$$\mu(\cdot|h_t^C, h_t^{P,i}) \in \Delta(\mathcal{X}^{-i}),\ \forall t. \tag{7}$$

For reasons that will become clear in the following, it is also helpful to track public beliefs

$$\mu(\cdot|h_t^C) \in \Delta(\boldsymbol{\mathcal{X}}),\ \forall t. \tag{8}$$

For a given correlation device ϕ^C, a belief system μ is consistent if $\mu(\cdot|h) = \mathbb{P}^{\phi^C}(\cdot|h)$ for any history h with $\mathbb{P}^{\phi^C}(h) > 0$. For h with $\mathbb{P}^{\phi^C}(h) = 0$, $\mu(\cdot|h)$ should be formed in a reasonable way and follow Bayes' rule if applicable. For the rest of this subsection, we will explain in details how a consistent belief system μ is formed in a recursive manner, for h with $\mathbb{P}^{\phi^C}(h) > 0$ or $\mathbb{P}^{\phi^C}(h) = 0$.

Prior to presenting the update rule for consistent beliefs, we present a useful property.

Lemma 1: At any time t, state components $x_{1:t}^i$ $(i = 1, \dots, N)$ are mutually independent conditioned on common observations h_t^C, i.e.,

$$\mathbb{P}^{\phi^C}(\boldsymbol{x}_{1:t}|\boldsymbol{a}_{1:t-1}, \boldsymbol{\gamma}_{1:t-1}) = \prod_{i=1}^{N} \mathbb{P}^{\phi^C}(x_{1:t}^i|\boldsymbol{a}_{1:t-1}, \boldsymbol{\gamma}_{1:t-1}). \tag{9}$$

Proof: See Appendix A. ∎

Lemma 1 plays an important role in belief formation. It indicates that all private beliefs can be formed through the public belief. By the mutual independence of x^i's conditioned on public observations, player i's private observation of $x_{1:t}^i$ plays no role in her belief formation toward x_t^{-i}, which implies that the private belief on x_t^{-i} should be identical to the common belief formed by public observations. Moreover, the actual prescription adopted by player i does not influence the belief on x_t^{-i}, no matter if she deviates or not. This means that even if player i is aware that the public belief is not completely correct owing to her own deviation, she can still utilize it to form private beliefs.

For a consistent belief system μ, based on Lemma 1, player i's private belief $\mu(\cdot|h_t^C, h_t^{P,i})$ for any information set $h_t^i = (h_t^C, h_t^{P,i})$ can be formed using only the public belief $\mu(\cdot|h_t^C)$ as

$$\mu_t(x_t^{-i}|h_t^C, h_t^{P,i}) = \prod_{j \neq i} \mu_t^j(x_t^j|h_t^C). \tag{10}$$

Lemma 2 derives a recursive formula for a consistent public belief system μ.

Lemma 2: Given the correlation device ϕ^C, if $\mathbb{P}^{\phi^C}(h_{t+1}^C) > 0$, a consistent belief system μ satisfies the following recursive equation

$$\mu_1(\boldsymbol{x}_1) = \prod_{i=1}^{N} \mu_1^i(x_1^i) = \prod_{i=1}^{N} Q^i(x_1^i), \tag{11}$$

$$\mu_{t+1}(\boldsymbol{x}_{t+1}|h_{t+1}^C) = \prod_{i=1}^{N} \mu_{t+1}^i(x_{t+1}^i|h_{t+1}^C), \tag{12}$$

with

$$\begin{aligned}
&\mu_{t+1}^i(x_{t+1}^i|h_{t+1}^C) \\
&= \frac{\sum_{x_t^i} Q^i(x_{t+1}^i|x_t^i, \boldsymbol{a}_t) 1_{\gamma_t^i(x_t^i)}(a_t^i) \mu_t^i(x_t^i|h_t^C)}{\sum_{x_t^i} 1_{\gamma_t^i(x_t^i)}(a_t^i) \mu_t^i(x_t^i|h_t^C)} \\
&=: \hat{T}\left(\mu_t^i(\cdot|h_t^C), \gamma_t^i, \boldsymbol{a}_t\right)(x_{t+1}^i),
\end{aligned} \tag{13}$$

which does not depend on ϕ^C.

Proof: See Appendix B. ∎

What if $\mathbb{P}^{\phi^C}(h_{t+1}^C) = 0$? If a common observation set h_{t+1}^C has zero measure under $\mathbb{P}^{\phi^C}$, it means at least one player i chose an a_τ^i that was not consistent with her received prescription γ_τ^i. In this case, every player knows player i deviated, but no one–other than player i–knows what was the exact adopted prescription. In this case, we can only make a reasonable speculation. In this work, we use the following (speculation): if player i has an obvious deviation (i.e., $a_t^i \neq \gamma_t^i(x_t^i)$ for any x_t^i with nonzero measure under the given belief), player i switched to the "constant" prescription $\tilde{\gamma}_t^i(x_t^i) \equiv a_t^i$. As a result, if a_t^i is an obvious deviation, μ_{t+1}^i is updated through

$$\mu_{t+1}^i(x_{t+1}^i|h_{t+1}^C) = \sum_{x_t^i} Q^i(x_{t+1}^i|x_t^i, \boldsymbol{a}_t) \mu_t^i(x_t^i|h_t^C), \tag{14}$$

which is consistent with (13) for constant prescriptions.

For the remaining part of the paper, we use the belief system μ defined by (13), (14) and denote by $\mathbb{P}^{\phi^C}$ the probability measure under the device ϕ^C and the belief system μ, with subscript μ omitted. Furthermore, in order to simplify the presentation, we define π_t as $\pi_t^i(\cdot) = \mu_t^i(\cdot|h_t^C)$, so that the conditioning on the public information is implied. The update of π_t follows the way described in Lemma 2 and (14) for μ_t.

B. Players' Rationality and Equilibrium Concept

Player i, at each time t chooses an action a_t^i based on her observation h_t^i. We use strategy sequence $g_{1:T}^i$ to describe player i's behavior, where $g_t^i : \mathcal{H}_t^i \mapsto \Delta(\mathcal{A}^i)$, so $A_t^i \sim g_t^i(\cdot|h_t^i)$. Specifically, an obedient strategy $g^{i,*}$ is defined as $g_t^{i,*}(\cdot|h_t^i) = 1_{\gamma_t^i(x_t^i)}(\cdot)$.

We define the reward-to-go function for player i as

$$W_t^{\phi^C,i}(h_t^i; g_{t:T}^i) = W_t^{\phi^C,i}(\boldsymbol{a}_{1:t-1}, \boldsymbol{\gamma}_{1:t-1}, x_{1:t}^i, \gamma_t^i; g_{t:T}^i)$$
$$= \mathbb{E}^{\phi^C, g_{t:T}^i}\left[\sum_{\tau=t}^{T} r_\tau^i(\boldsymbol{X}_\tau, \boldsymbol{A}_\tau)|h_t^i\right], \quad (15)$$

based on which we define the perfect correlated equilibrium.

Definition 1: A correlation device ϕ^C is said to be a PCE if for every player i, any observation h_t^i with γ_t^i such that $\hat{\phi}_t^C(\gamma_t^i|h_t^C) > 0$, and any g^i:

$$W_t^{\phi^C,i}(h_t^i; g_{t:T}^{i,*}) \geq W_t^{\phi^C,i}(h_t^i; g_{t:T}^i). \quad (16)$$

IV. Structured Equilibria

In practice, to verify or design a PCE ϕ^C through Definition 1, one will need to evaluate the conditional expectations (15) at each time t, for each player i, with all possible private histories h_t^i and actions a_t^i. As the dimension of h_t^i increases with time t, complexity of the forms of both constraints as well as the device ϕ_t^C grows. This motivates our investigation on PCE with a special structure. In this section, we propose the concept of structured correlation device. A structured correlation device $\hat{\phi}^C$ has the form

$$\hat{\phi}_t^C : \Delta(\boldsymbol{\mathcal{X}}) \mapsto \Delta(\boldsymbol{\mathcal{G}}). \quad (17)$$

A PCE with structured correlation device is called a structured PCE (sPCE).

A. Motivation for Structured Devices

Structured devices simplify the design of correlation device by summarizing common histories h_t^C with public beliefs π_t. This simplification may raise the following concerns: (i) does this reduction influence the achievable outcomes? and (ii) given this reduction, is it sufficient for rational players to check structured deviated strategies?

Theorem 1 shows that all the expected payoff profile induced by general devices can be achieved by structured ones.

Theorem 1 (Expected Payoff Preservation): For any correlation device ϕ^C, one can find a structured device $\hat{\phi}^C$, such that for all players $i = 1, \ldots, N$, the expected payoffs remain the same, i.e., $\forall i, t$

$$\mathbb{E}^{\phi^C}\left[r_t^i(\boldsymbol{X}_t, \boldsymbol{A}_t)\right] = \mathbb{E}^{\hat{\phi}^C}\left[r_t^i(\boldsymbol{X}_t, \boldsymbol{A}_t)\right]. \quad (18)$$

Proof: See Appendix C. ∎

The above theorem provides sufficient motivation to focus on structured devices. Next, we explore the structure of players' rationality given a structured correlation device.

Theorem 2: When a structured device $\hat{\phi}^C$ is used, a rational player i is faced with a Markov decision process (MDP) with state $(\pi_t, x_t^i, \gamma_t^i)$, action a_t^i, and instantaneous reward

$$\mathbb{E}^{\hat{\phi}^C, g^i}\left[r_t^i(\boldsymbol{X}_t, \boldsymbol{A}_t)|h_t^i, a_t^i\right]$$
$$= \mathbb{E}\left[r_t^i(\boldsymbol{X}_t, \boldsymbol{A}_t)|\pi_t, x_t^i, \psi_t, \gamma_t^i, a_t^i\right]$$
$$=: \bar{r}_t^i(\pi_t, x_t^i, \gamma_t^i, a_t^i; \psi_t), \quad (19)$$

where we define ψ_t as the prescription profile distribution given by $\hat{\phi}_t^C$ under π_t, i.e., $\psi_t := \hat{\phi}_t^C[\pi_t]$.

Proof: See Appendix D. ∎

Theorem 2 provides an important motivation for focusing on structured devices. Although rational players may consider arbitrary unilateral deviated strategies that depend on the complete observation $h_t^i = (h_t^C, h_t^{i,P})$, this theorem indicates that it is sufficient for them to only check deviations that depend on h_t^C through π_t. This gives further justification to the choice of "structured" devices in (17).

From known results on MDPs [16, Chap. 6], due to Theorem 2, it is sufficient for player i to check deviations that are Markov deterministic strategies $\hat{g}^i$, i.e., strategies that generate a_t^i as $a_t^i = \hat{g}_t^i(\pi_t, x_t^i, \gamma_t^i)$. Therefore, with a structured correlation device, players don't have to track the complete h_t^i but only π_t, x_t^i, γ_t^i for the purpose of verifying players' rationality.

B. Backward Recursive Characterization for sPCE

Inspired by Theorem 2, in this subsection we describe a backward recursive characterization for sPCE.

Theorem 3 (Backward Recursive Characterization): A structured device $\hat{\phi}^C$ is an sPCE if and only if it passes the verification steps described as follows:

1) For all i, set $\bar{V}_{T+1}^i(\pi_{T+1}, x_{T+1}^i) \equiv 0$.
2) Start from $t = T$, for all i,
   a) Verify the inequalities for all π_t, $\psi_t = \hat{\phi}_t^C[\pi_t]$, for all γ_t^i with $\psi_t(\gamma_t^i) > 0$, and all x_t^i, a_t^i,

$$\bar{r}_t^i(\pi_t, x_t^i, \gamma_t^i, \gamma_t^i(x_t^i); \psi_t)$$
$$+ \mathbb{E}[\bar{V}_{t+1}^i(\Pi_{t+1}, X_{t+1}^i)|\pi_t, x_t^i, \gamma_t^i, \gamma_t^i(x_t^i), \psi_t]$$
$$\geq \bar{r}_t^i(\pi_t, x_t^i, \gamma_t^i, a_t^i; \psi_t)$$
$$+ \mathbb{E}[\bar{V}_{t+1}^i(\Pi_{t+1}, X_{t+1}^i)|\pi_t, x_t^i, \gamma_t^i, a_t^i, \psi_t]. \quad (20)$$

   $\hat{\phi}^C$ does not pass the verification test if one of the inequalities does not hold.
   b) For all π_t, $\psi_t = \hat{\phi}_t^C[\pi_t]$, for all γ_t^i with $\psi_t(\gamma_t^i) > 0$, and all x_t^i, update the quantities

$$\bar{V}_t^i(\pi_t, x_t^i) = \mathbb{E}[\bar{r}_t^i(\Pi_t, X_t^i, \Gamma_t^i, \Gamma_t^i(X_t^i); \Psi_t)$$
$$+ \bar{V}_{t+1}^i(\Pi_{t+1}, X_{t+1}^i)|\pi_t, x_t^i, \psi_t]. \quad (21)$$

   c) The verification completes if $t = 1$. Otherwise, set $t \leftarrow t - 1$ and go back to (a).

Proof: See Appendix E. ∎

The verification steps in Theorem 3 have a much smaller computational complexity compared to the brute force method suggested in (16). Assuming identical agents, for a single time step t, the number of constraints in Theorem 3 is $N|\mathcal{P}_t||\mathcal{G}^i||\mathcal{X}^i||\mathcal{A}^i|$, while the number of constraints in (16) is $N|\mathcal{H}_t^i|(\#g_{t:T}^i)$, where $|\mathcal{H}_t^i| = |\mathcal{H}_t^c||\mathcal{X}^i|^t|\mathcal{G}^i|$ is larger than $|\mathcal{P}_t||\mathcal{X}^i||\mathcal{G}^i|$, and $(\#g_{t:T}^i)$ is the number of possible current and future deviation strategies of user i and grows double exponentially with t.

Indeed, Theorem 3 offers more than a systematic approach to check if a structured device $\hat{\phi}^C$ is a PCE. Notice that up to the verification of time step t, only $\hat{\phi}_{t:T}^C$ are involved, and once $\bar{V}_t^i$ is formed, $\hat{\phi}_t^C$ is no longer needed in computation. Based on these observations, one may also construct an

sPCE following the steps described in Theorem 3, with some necessary modifications. Specifically, in step 2a, we construct a $\hat{\phi}_t^C$ using the following greedy selection: for each π_t, find a ψ_t, such that (20) holds, and set $\hat{\phi}_t^C[\pi_t] = \psi_t$. Although this greedy approach does not guarantee a nonempty candidate set for ψ_t in each time step t and belief π_t, if one successfully constructs $\hat{\phi}_{1:T}^C$ by this approach, Theorem 3 guarantees that the result $\hat{\phi}^C$ is an sPCE.

C. The Connection between sPCE and sPBE

The structured perfect Bayesian equilibrium (sPBE) introduced in [5] is a solution concept defined also on Markov games with asymmetric information, with a similar form as the sPCE discussed here. Without coordination from a third party, at each time t each player i determines her own (randomized) prescription $\gamma_t^i : \mathcal{X}^i \mapsto \Delta(\mathcal{A}^i)$ with $\gamma_t^i = \theta_t^i(\pi_i)$ through the equilibrium strategy θ_t^i, where π_t is a belief vector whose i-th component π_t^i is a belief on x_t^i conditioned on public observation $\boldsymbol{a}_{1:t-1}$ and the common knowledge $\boldsymbol{\gamma}_{1:t-1}$ in equilibrium with $\gamma_\tau^i = \theta_\tau^i(\pi_\tau)$. Due to the similar structures of sPCE and sPBE, the outcome of a subset of sPCE can be realized by sPBE.

Lemma 3: An sPCE $\hat{\phi}^C$ is realizable by an sPBE if and only if $\hat{\phi}_t^C[\pi_t]$ assigns probability 1 to a single prescription profile $\boldsymbol{\gamma}_t$ for all t and π_t (i.e., $\hat{\phi}^C$ is deterministic). Here the realizability means (i) the same public information $h_t^C = (\boldsymbol{a}_{1:t-1}, \boldsymbol{\gamma}_{1:t-1})$ leads to the same belief π_t; (ii) the same belief π_t induces the same distribution on $\boldsymbol{\gamma}_t$.

Proof: See Appendix F. ∎

One may doubt the necessity of the deterministic correlation device for realizability by sPBE, since in static games, a correlated equilibrium can be realized by a Nash equilibrium as long as the distribution on strategy profiles shows mutual independence among strategies of different players. However, mutual independence is not sufficient for an sPCE to be realized by sPBE. This insufficiency is caused by the different information structure under our coordination settings. The broadcast of $\boldsymbol{\gamma}_{t-1}$ at time t provides extra information for state inference, which is absent in the sPBE setting. As a result, even if one uses some "sPBE" with randomized prescriptions to simulate an identical distribution of actions as that of an sPCE with mutual independence, due to the lack of the communication channel for $\boldsymbol{\gamma}$, this "sPBE" is not able to induce a belief on the state as informative as that induced by the sPCE in general.

V. A Concrete Example: Public Investment Game

In this section, we demonstrate how the backward dynamic approach described in Theorem 3 is used in the design of sPCE through an example with a public investment game, which was also used in [5].

The public investment problem that we consider in this paper is a two-stage dynamic game with two players. The task of the two players is to decide whether to invest at each stage. In the beginning of the game, each player i privately knows her cost type $x^i \in \{L, H\}$ for the investment (with the meaning Low and High cost, $L < 1 < H$), where the type $\boldsymbol{X} = (X^1, X^2)$ is considered to be drawn from a commonly known distribution $Q^i(X^i = H) = q,\ i = 1, 2$. At each stage t, each player i decides if she invests the public good ($a_t^i = 1$, with a cost x^i) or not ($a_t^i = 0$, with no cost). If any one of them invests at stage t, both players receive 1 unit of benefit. If no one invests, both players earn nothing. This setting captures the effect of "free riding". The social welfare can improve if only one player invests at each stage (preferably the one with the lowest cost, if one exists). Thus, coordination between players is very desirable. The players' instantaneous rewards can be expressed as

$$r_t^i(\boldsymbol{x}, \boldsymbol{a}_t) = 1_1(a_t^i) \cdot (1 - x^i) + 1_0(a_t^i) \cdot 1_1(a_t^{-i}). \tag{22}$$

We use $\pi_t = (\pi_t^1, \pi_t^2)$ to denote the public belief on $\boldsymbol{x}$ at stage t, where π_t^i is the probability of $X^i = H$ given the public observation up to time t. Therefore, the initial public belief is $\boldsymbol{\pi}_1 = (q, q)$. We want to design an sPCE, which specifies probability distributions over the prescription profile $\boldsymbol{\gamma}_t = (\gamma_t^1, \gamma_t^2) \in \mathcal{G}^1 \times \mathcal{G}^2$ at time t given π_t, where

$$\gamma_t^i \in \mathcal{G}^i = \{\gamma_{00}, \gamma_{01}, \gamma_{10}, \gamma_{11}\}, \tag{23}$$

and γ_{mn} with $m, n \in \{0, 1\}$ means "play m if $x^i = L$, and play n if $x^i = H$". Since $\boldsymbol{\mathcal{G}}$ and $\mathcal{A}^i$ are finite set, $\pi_1^i = q$, the set $\mathcal{P}_2$ of all possible π_2^i is also finite because $\pi_2^i = \hat{T}(\pi_1^i, \gamma_1^i, a_1^i)$. We have $\mathcal{P}_1 = \{q\}$, $\mathcal{P}_2 = \{0, q, 1\}$, and $\pi_t \in (\mathcal{P}_t)^2$. Intuitively, if one plays γ_{01} or γ_{10} she perfectly reveals her true type, while playing γ_{00} or γ_{11} does not change the public belief about her true type.

To find an sPCE, we utilize the backward dynamic approach described in Theorem 3. We set $V_3^i \equiv 0$. At each stage t, if V_{t+1}^i's are given, for each $\pi_t \in (\mathcal{P}_t)^2$, the design of $\hat{\phi}_t^C[\pi_t]$ is essentially finding a point ψ_t that satisfies linear constraints (20). For certain π_t's, there can be infinitely many ψ_t's satisfying the constraints, so here we select the expected social welfare to go (i.e., the sum of players' reward from time t to T) conditioned on the public observation up to time t as the objective function to maximize, so that we can use linear programming to find a solution (also as a greedy approach to maximize the expected social welfare). Once $\hat{\phi}_t^C[\pi_t]$ is found, we follow (21) to evaluate V_t^i, and pass it to the design task at stage $(t-1)$. We implement the approach described above numerically, with $L = 0.2, H = 1.2$, for q ranging from 0.01 to 0.99 with step 0.01. We also solve the corresponding team problem, where agents are non-strategic and act as a team trying to maximize the social welfare, using the common information approach described in [17].

Figure 1 shows a plot of the result. The sPCE are found at q values from 0.01 to 0.20, and from 0.34 to 0.99. For the remaining values of $q \in [0.21, 0.33]$ our greedy algorithm did not find any sPCE. The expected social welfare provided by sPCE is close to the optimal welfare in the team problem for small q (from 0.01 to 0.20), but the gap between these two becomes large when q is large. This is not surprising, because in the team problem, even if two players are both of H type, choosing exactly one of them to invest is a profitable move for the society, but this violates players' rationality in the game problem. Thus, a larger q leads to a smaller

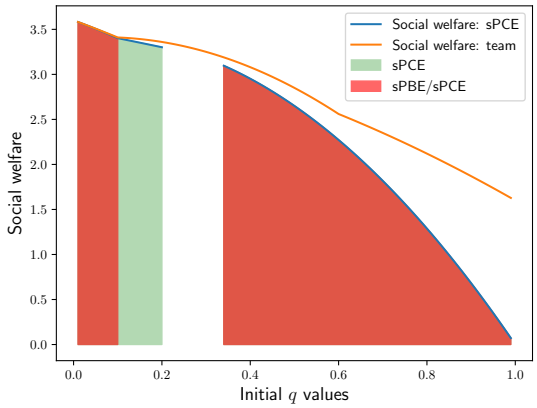

Fig. 1. The expected social welfare versus q values. Note that all the solutions marked as sPBE are also sPCE. The blank space between $q = 0.21$ to $q = 0.33$ indicates no solution is found by our experiment algorithm, but this does not necessarily imply sPCE does not exist.

probability of investment in the game, resulting in smaller social welfare.

To get a sense of what an sPCE looks like, we list the details of the correlation device for $q = 0.2$ in Table I. For each π_2 at stage 2, the prescription profiles proposed by the correlation device are indeed pure Nash equilibria of the stage game. At stage 1, the correlation device will always tell player 2 to invest if $x^2 = L$, and not to invest if $x^2 = H$. For player 1, with probability $5/6$ she will be told the same suggestion as player 2, but with probability $1/6$ she will be told not to invest in any case. As an example, let's justify the rationality of player 1 with type L and suggested prescription γ_{10}. Since both players' prescriptions are γ_{10} in this case, by the belief update rule, the action a_1^i will act as an announcement of player i's type. Suppose player 1 follows the suggestion to invest at stage 1, she earns instantaneous reward 0.8 no matter what player 2 does. Then, with probability 0.2 the next $\pi_2 = (0,0)$, in which case player 1 can fully rely on player 2's investment and earn 1 for sure; or with probability 0.8 the next $\pi_2 = (0,1)$, in which she knows player 2 won't invest and thus she invests to obtain 0.8. The expected reward turns out to be $0.8+0.8\times 1+0.2\times 0.8 = 1.76$. However, if player 1 chooses to disguise herself as a type H by playing $a_1^1 = 0$, then with probability 0.2, player 2 is of type H so that $\pi_2 = (1,1)$, in which case player 1 earns nothing; or with probability 0.8, player 2 is of type L, so that $\pi_2 = (1,0)$, and player 1 earns 1 in the second stage. The total expectation of reward for this deviation is $0.8 \times 1 = 0.8$, so player 1 won't deviate.

VI. Conclusion

We studied coordination in Markov games with asymmetric information through a new solution concept, namely PCE. Motivated by the sufficiency of structured devices in terms of payoff profiles, together with the MDP characterization of players' rationality, we proposed a sPCE, and developed a backward dynamic approach that works for both verification and design of sPCE. Through a numerical experiment, we demonstrated how the backward dynamic approach works.

TABLE I

An sPCE device $\hat{\phi}^C$ for $q = 0.2$.

$(V_t^i(\pi_t,\cdot) = (V_t^i(\pi_t, L), V_t^i(\pi_t, H)))$

t	(π_t^1, π_t^2)	$\hat{\phi}_t^C(\cdot\|\pi_t)$			$V_t^1(\pi_t,\cdot)$	$V_t^2(\pi_t,\cdot)$
		γ_{00},γ_{10}	γ_{10},γ_{00}	γ_{10},γ_{10}		
1	$(0.2, 0.2)$	$\mathbf{1/6}$	0	$\mathbf{5/6}$	$(1.76, 1.6)$	$(1.6, 1.47)$
2	$(0,0)$	$\mathbf{1}$	0	0	$(1,1)$	$(0.8,0)$
	$(0,0.2)$	0	$\mathbf{1}$	0	$(0.8,0)$	$(1,1)$
	$(0,1)$	0	$\mathbf{1}$	0	$(0.8,0)$	$(1,1)$
	$(0.2,0)$	1	0	0	$(1,1)$	$(0.8,0)$
	$(0.2,0.2)$	0	0	$\mathbf{1}$	$(0.8,0.8)$	$(0.8,0.8)$
	$(0.2,1)$	0	0	$\mathbf{1}$	$(0.8,0)$	$(0.8,0.8)$
	$(1,0)$	$\mathbf{1}$	0	0	$(1,1)$	$(0.8,0)$
	$(1,0.2)$	$\mathbf{1}$	0	0	$(0.8,0.8)$	$(0.8,0)$
	$(1,1)$	0	0	$\mathbf{1}$	$(0.8,0)$	$(0.8,0)$

As a future research direction, the idea of PCE in this paper may be extended to more general decentralized MDP scenarios (MDP with infinite time horizon, x_t^i not directly observable by i, etc.) with strategic agents, as a new option for coordination. It is also worth noting that the algorithm used in the numerical experiment is heuristic and does not guarantee a solution, so it is possible to develop a more effective algorithm based on our backward dynamic approach. Possible techniques such as constrained MDP [18] and online optimization with hard constraints [19] can be used to develop learning algorithms for sPCE, or sPCE with some ϵ-relaxation in rationality constraints. Potential future directions can also involve the structural results of sPCE in mean-field games, and combining the coordinator idea with mechanism design.

Appendix

A. Proof of Lemma 1

Proof: Consider the joint distribution of $\boldsymbol{x}_{1:t}$, $\boldsymbol{a}_{1:t-1}$ and $\boldsymbol{\gamma}_{1:t-1}$,

$$\begin{aligned}&\mathbb{P}^{\phi^C}(\boldsymbol{x}_{1:t}, \boldsymbol{a}_{1:t-1}, \boldsymbol{\gamma}_{1:t-1})\\ &= \prod_{i=1}^N Q^i(x_1^i) \prod_{\tau=1}^{t-1} 1_{\gamma_\tau^i(x_\tau^i)}(a_\tau^i)\\ &\quad \cdot \phi_\tau^C(\boldsymbol{\gamma}_\tau|\boldsymbol{a}_{1:\tau-1}, \boldsymbol{\gamma}_{1:\tau-1}) Q^i(x_{\tau+1}^i|x_\tau^i, a_\tau). \end{aligned} \tag{24}$$

Therefore,

$$\begin{aligned}&\mathbb{P}^{\phi^C}(\boldsymbol{x}_{1:t}|\boldsymbol{a}_{1:t-1}, \boldsymbol{\gamma}_{1:t-1})\\ &= \frac{\mathbb{P}^{\phi^C}(\boldsymbol{x}_{1:t}, \boldsymbol{a}_{1:t-1}, \boldsymbol{\gamma}_{1:t-1})}{\sum_{\bar{\boldsymbol{x}}_{1:t}} \mathbb{P}^{\phi^C}(\bar{\boldsymbol{x}}_{1:t}, \boldsymbol{a}_{1:t-1}, \boldsymbol{\gamma}_{1:t-1})} = \frac{NUM}{DEN}, \end{aligned} \tag{25}$$

where

$$NUM = \prod_{i=1}^N Q^i(x_1^i) \prod_{\tau=1}^{t-1} 1_{\gamma_\tau^i(x_\tau^i)}(a_\tau^i) Q^i(x_{\tau+1}^i|x_\tau^i, \boldsymbol{a}_\tau),$$

and DEN is nothing but substituting x in NUM with $\bar{x}$ and then do the summation over $\bar{x}_{1:t}$. Note that the ϕ_τ^C-terms disappear because both NUM and DEN have them, so they are cancelled out. Then, one may move the product over player indices to the front of the fraction, as

$$\begin{aligned}&\mathbb{P}^{\phi^C}(\boldsymbol{x}_{1:t}|\boldsymbol{a}_{1:t-1}, \boldsymbol{\gamma}_{1:t-1})\\ &= \prod_{i=1}^N \frac{Q^i(x_1^i) \prod_{\tau=1}^{t-1} 1_{\gamma_\tau^i(x_\tau^i)}(a_\tau^i) Q^i(x_{\tau+1}^i|x_\tau^i, \boldsymbol{a}_\tau)}{\sum_{\bar{x}_{1:t}} Q^i(\bar{x}_1^i) \prod_{\tau=1}^{t-1} 1_{\gamma_\tau^i(\bar{x}_\tau^i)}(a_\tau^i) Q^i(\bar{x}_{\tau+1}^i|\bar{x}_\tau^i, \boldsymbol{a}_\tau)}\\ &= \prod_{i=1}^N \mathbb{P}^{\phi^C}(x_{1:t}^i|\boldsymbol{a}_{1:t-1}, \boldsymbol{\gamma}_{1:t-1}). \end{aligned} \tag{26}$$

■

B. Proof of Lemma 2

Proof: The equation for $\mu_1(\boldsymbol{x}_1)$ is straightforward. Suppose the recursive equation is true up to time t. For time $t+1$, if $\mathbb{P}^{\phi^C}(h_{t+1}^C) > 0$,

$$\mu_{t+1}(\boldsymbol{x}_{t+1}|h_{t+1}^C) = \mathbb{P}^{\phi^C}(\boldsymbol{x}_{t+1}|\boldsymbol{\gamma}_{1:t}, \boldsymbol{a}_{1:t}) = \frac{\sum_{\boldsymbol{x}_t} \mathbb{P}^{\phi^C}(\boldsymbol{x}_{t+1}, \boldsymbol{x}_t, \boldsymbol{a}_t, \boldsymbol{\gamma}_t|\boldsymbol{\gamma}_{1:t-1}, \boldsymbol{a}_{1:t-1})}{\sum_{\boldsymbol{x}_t} \mathbb{P}^{\phi^C}(\boldsymbol{x}_t, \boldsymbol{a}_t, \boldsymbol{\gamma}_t|\boldsymbol{\gamma}_{1:t-1}, \boldsymbol{a}_{1:t-1})}. \tag{27}$$

$$\begin{aligned}&\mathbb{P}^{\phi^C}(\boldsymbol{x}_{t+1}, \boldsymbol{x}_t, \boldsymbol{a}_t, \boldsymbol{\gamma}_t|\boldsymbol{\gamma}_{1:t-1}, \boldsymbol{a}_{1:t-1}) \\ &= \prod_{i=1}^N Q^i(x_{t+1}^i|x_t^i, \boldsymbol{a}_t) 1_{\gamma_t^i}(a_t^i) \\ &\quad \cdot \phi_t^C(\boldsymbol{\gamma}_t|\boldsymbol{\gamma}_{1:t-1}, \boldsymbol{a}_{1:t-1}) \mu_t(\boldsymbol{x}_t|h_t^C).\end{aligned} \tag{28}$$

Substituting this it back to (27), and recalling that $\mu_t(\boldsymbol{x}_t|h_t^C)$ can be written as the product of $\mu_t^i(x_t^i|h_y^C)$ by induction assumption we have

$$\begin{aligned}&\mu_{t+1}(\boldsymbol{x}_{t+1}|h_{t+1}^C) \\ &= \frac{\sum_{\boldsymbol{x}_t} \prod_{i=1}^N Q^i(x_{t+1}^i|x_t^i, \boldsymbol{a}_t) 1_{\gamma_t^i(x_t^i)}(a_t^i) \mu_t^i(x_t^i|h_t^C)}{\sum_{\boldsymbol{x}_t} \prod_{i=1}^N 1_{\gamma_t^i(x_t^i)}(a_t^i) \mu_t^i(x_t^i|h_t^C)} \\ &= \prod_{i=1}^N \frac{\sum_{x_t^i} Q^i(x_{t+1}^i|x_t^i, \boldsymbol{a}_t) 1_{\gamma_t^i(x_t^i)}(a_t^i) \mu_t^i(x_t^i|h_t^C)}{\sum_{x_t^i} 1_{\gamma_t^i(x_t^i)}(a_t^i) \mu_t^i(x_t^i|h_t^C)} \\ &=: \prod_{i=1}^N \mu_{t+1}^i(x_{t+1}^i|h_{t+1}^C),\end{aligned} \tag{29}$$

where for the first equality, ϕ_t^C-terms are cancelled. ■

C. Proof of Theorem 1

Proof: We prove it through the following lemma.

Lemma 4: Given general device ϕ^C, one can find a structured device $\hat{\phi}^C$, such that for all t,

$$\mathbb{P}^{\hat{\phi}^C}(\pi_t, \gamma_t, \pi_{t+1}) = \mathbb{P}^{\phi^C}(\pi_t, \gamma_t, \pi_{t+1}). \tag{30}$$

We prove Theorem 1 using Lemma 4. For any device ϕ^C,

$$\mathbb{P}^{\phi^C}(\boldsymbol{x}_t, \boldsymbol{a}_t|\pi_t, \boldsymbol{\gamma}_t) = \pi_t(\boldsymbol{x}_t) \prod_{i=1}^N 1_{\gamma_t^i(x_t^i)}(a_t^i), \tag{31}$$

which does not depend on ϕ^C. Accordingly,

$$\tilde{r}_t^i(\pi_t, \boldsymbol{\gamma}_t) := \mathbb{E}^{\phi^C}[r_t^i(\boldsymbol{X}_t, \boldsymbol{A}_t)|\pi_t, \boldsymbol{\gamma}_t] \tag{32}$$

does not depend on ϕ^C. By law of iterated expectation,

$$\mathbb{E}^{\phi^C}\left[r_t^i(\boldsymbol{X}_t, \boldsymbol{A}_t)\right] = \mathbb{E}^{\phi^C}\left[\tilde{r}_t^i(\Pi_t, \boldsymbol{\Gamma}_t)\right].$$

From Lemma 4, we can construct a $\hat{\phi}^C$, such that the joint distribution on $(\Pi_t, \boldsymbol{\Gamma}_t)$ are the same for every t, which implies the same expected total rewards under ϕ^C and $\hat{\phi}^C$ for every player i, because each expectation term on the right hand side of (33) depends only on the joint distribution.

Proof of Lemma 4. The proof is done by forward induction. For general device ϕ^C, at time t, the joint distribution on $(\pi_t, \boldsymbol{\gamma}_t, \pi_{t+1})$ depends on ϕ^C through $\phi_{1:t}^C$ because these variables are functions of h_t^C. For $t = 1$, we simply choose $\hat{\phi}_1^C = \phi_1^C$, and (30) holds. For time t, suppose

$$\mathbb{P}^{\hat{\phi}_{1:t-1}^C}(\pi_{t-1}, \gamma_{t-1}, \pi_t) = \mathbb{P}^{\phi_{1:t-1}^C}(\pi_{t-1}, \gamma_{t-1}, \pi_t). \tag{33}$$

Construct $\hat{\phi}_t^C$ such that

$$\begin{aligned}\hat{\phi}_t^C(\boldsymbol{\gamma}_t|\pi_t) = \mathbb{P}^{\phi^C}(\boldsymbol{\gamma}_t|\pi_t) &= \frac{\mathbb{P}^{\phi^C}(\pi_t, \boldsymbol{\gamma}_t)}{\mathbb{P}^{\phi_{1:t-1}^C}(\pi_t)} \\ &= \frac{\sum_{h_t^C:\pi_t} \mathbb{P}^{\phi_{1:t-1}^C}(h_t^C) \phi_t^C(\boldsymbol{\gamma}_t|h_t^C)}{\mathbb{P}^{\phi_{1:t-1}^C}(\pi_t)},\end{aligned} \tag{34}$$

where $h_t^C : \pi_t$ means the set of h_t^C that induces π_t. Thus,

$$\begin{aligned}&\mathbb{P}^{\hat{\phi}^C}(\pi_t, \boldsymbol{\gamma}_t, \pi_{t+1}) \\ &= \sum_{\boldsymbol{x}_t, \boldsymbol{a}_t} \mathbb{P}^{\hat{\phi}^C}(\pi_t) \pi_t(\boldsymbol{x}_t) \hat{\phi}_t^C(\boldsymbol{\gamma}_t|\pi_t) \\ &\quad \cdot \prod_{i=1}^N 1_{\gamma_t^i(x_t^i)}(a_t^i) 1_{\hat{T}(\pi_t, \boldsymbol{\gamma}_t, \boldsymbol{a}_t)}(\pi_{t+1}) \\ &= \sum_{\boldsymbol{x}_t, \boldsymbol{a}_t} \mathbb{P}^{\hat{\phi}^C}(\pi_t) \pi_t(\boldsymbol{x}_t) \mathbb{P}^{\phi^C}(\boldsymbol{\gamma}_t|\pi_t) \\ &\quad \cdot \prod_{i=1}^N 1_{\gamma_t^i(x_t^i)}(a_t^i) 1_{\hat{T}(\pi_t, \boldsymbol{\gamma}_t, \boldsymbol{a}_t)}(\pi_{t+1}) = \mathbb{P}^{\phi^C}(\pi_t, \boldsymbol{\gamma}_t, \pi_{t+1}).\end{aligned} \tag{35}$$

By induction, Lemma 4 holds. ■

D. Proof of Theorem 2

Proof: State transition. This part shows π_t, x_t^i, γ_t^i are sufficient statistics of h_t^i with respect to $\pi_{t+1}, x_{t+1}^i, \gamma_{t+1}^i$, and the belief formation does not depend on g^i. Suppose player i uses strategy g^i such that at each time t it generates a_t^i based on h_t^i,

$$\begin{aligned}&\mathbb{P}^{\hat{\phi}^C, g^i}(\pi_{t+1}, x_{t+1}^i, \gamma_{t+1}^i|h_t^i, a_t^i) \\ &= \sum_{x_t^{-i}, a_t^{-i}} \pi_t(x_t^{-i}) \sum_{\gamma_t^{-i}} \hat{\phi}_t^C(\gamma_t^{-i}|\pi_t, \gamma_t^i) \prod_{j \neq i} 1_{\gamma_t^j(x_t^j)}(a_t^j) \\ &\quad \cdot 1_{\hat{T}(\pi_t, \boldsymbol{\gamma}_t, \boldsymbol{a}_t)}(\pi_{t+1}) Q^i(x_{t+1}^i|x_t^i, \boldsymbol{a}_t) \hat{\phi}_{t+1}^C(\gamma_t^i|\pi_{t+1}) \\ &= \mathbb{P}^{\hat{\phi}^C}(\pi_{t+1}, x_{t+1}^i, \gamma_{t+1}^i|\pi_t, x_t^i, \gamma_t^i, a_t^i),\end{aligned} \tag{36}$$

where π_t can be obtained from h_t^C (which is a part of h_t^i) according to the update rule $\hat{T}$.

Instantaneous reward. The expectation depends on the following probability

$$\begin{aligned}&\mathbb{P}^{\hat{\phi}^C, g^i}(\tilde{\boldsymbol{x}}_t, \tilde{\boldsymbol{a}}_t|h_t^i, a_t^i) \\ &= 1_{x_t^i}(\tilde{x}_t^i) \pi_t(\tilde{x}_t^{-i}) \sum_{\gamma_t^{-i}} \psi_t(\gamma_t^{-i}|\gamma_t^i) \prod_{j \neq i} 1_{\gamma_t^j(x_t^j)}(\tilde{a}_t^j) \cdot 1_{a_t^i}(\tilde{a}_t^i) \\ &= \mathbb{P}(\tilde{\boldsymbol{x}}_t, \tilde{\boldsymbol{a}}_t|\pi_t, x_t^i, \gamma_t^i, a_t^i, \psi_t),\end{aligned} \tag{37}$$

which implies that $\mathbb{E}^{\hat{\phi}^C, g^i}[r_t^i(\boldsymbol{X}_t, \boldsymbol{A}_t)|h_t^i, \psi_{1:t}, a_t^i]$ only depends on state $(\pi_t, x_t^i, \gamma_t^i)$, action a_t^i, and $\psi_t = \hat{\phi}^C[\pi_t]$. Therefore, $\bar{r}_t^i(\pi_t, x_t^i, \gamma_t^i, a_t^i; \psi_t)$ is a valid instantaneous reward for MDP given ψ_t is determined by $\hat{\phi}^C, \pi_t$.

Therefore, given $\hat{\phi}^C$, player i is faced with an MDP. ■

E. Proof of Theorem 3

Proof: According to Definition 1 and Theorem 2, $\hat{\phi}^C$ is an sPCE if and only if the obedient strategy $g^{i,*}$ is the optimal strategy for each player's MDP problem. Note that $g^{i,*}$ is a Markov deterministic strategy for player i's MDP, given $\hat{\phi}^C$, the value function satisfies the Bellman equations:

$$V^i_{T+1}(\pi_{T+1}, x^i_{T+1}, \gamma^i_{T+1}; \psi_t) \equiv 0, \tag{38a}$$

$$J^i_t(\pi_t, x^i_t, \gamma^i_t, a^i_t; \psi_t) := \bar{r}^i_t(\pi_t, x^i_t, \gamma^i_t, a^i_t; \psi_t) + \mathbb{E}^{\hat{\phi}^C}[V^i_{t+1}(\Pi_{t+1}, X^i_{t+1}, \Gamma^i_{t+1}; \Psi_{t+1})|\pi_t, x^i_t, \gamma^i_t, a^i_t, \psi_t], \tag{38b}$$

$$V^i_t(\pi_t, x^i_t, \gamma^i_t; \psi_t) := J^i_t(\pi_t, x^i_t, \gamma^i_t, \gamma^i_t(x^i_t); \psi_t), \tag{38c}$$

where the ψ_t is determined by $\hat{\phi}^C_t[\pi_t]$ uniquely, but we put it here to emphasize that the dependency on $\hat{\phi}^C_t$ is only through $\psi_t = \hat{\phi}^C_t[\pi_t]$. The rationality of the players are equivalent to the optimality of $g^{i,*}$ for players' MDP, which can be written as: $\forall t, \pi_t$, $\psi_t = \hat{\phi}^C_t[\pi_t]$, if $\psi_t(\gamma^i_t) > 0$, then for all a^i_t,

$$V^i_t(\pi_t, x^i_t, \gamma^i_t; \psi_t) \geq J^i_t(\pi_t, x^i_t, \gamma^i_t, a^i_t; \psi_t). \tag{39}$$

For the details of Bellman equation and the optimality condition, see [16, Chap. 6, Thm. 2.15].

The Bellman equations (38) and the optimality condition (39) provide sufficient and necessary conditions of sPCE for a structured device $\hat{\phi}^C$. Nevertheless, the above is not a complete time decomposition, since, in order to evaluate the expectation in (38b), we need to know $\hat{\phi}^C_{t+1}$. To resolve this issue, instead of tracking V, we define the following object:

$$\bar{V}^i_t(\pi_t, x^i_t) := \mathbb{E}^{\hat{\phi}^C}[V^i_t(\Pi_t, X^i_t, \Gamma^i_t; \Psi_t)|\pi_t, x^i_t], \tag{40}$$

with which, by law of iterated expectation, in (38b), we have

$$\begin{aligned}
&\mathbb{E}^{\hat{\phi}^C}[V^i_{t+1}(\Pi_{t+1}, X^i_{t+1}, \Gamma^i_{t+1}; \Psi_{t+1})|\pi_t, x^i_t, \gamma^i_t, a^i_t, \psi_t] \\
&= \mathbb{E}^{\hat{\phi}^C}[\mathbb{E}^{\hat{\phi}^C}[V^i_{t+1}(\Pi_{t+1}, X^i_{t+1}, \Gamma^i_{t+1}; \Psi_{t+1})|\Pi_{t+1}, X^i_{t+1}] \\
&\quad |\pi_t, x^i_t, \gamma^i_t, a^i_t, \psi_t] \\
&= \mathbb{E}[\bar{V}^i_{t+1}(\Pi_{t+1}, X^i_{t+1})|\pi_t, x^i_t, \gamma^i_t, a^i_t, \psi_t],
\end{aligned} \tag{41}$$

where the dependency on $\hat{\phi}^C$ is dropped because

$$\begin{aligned}
&\mathbb{P}^{\hat{\phi}^C}(\pi_{t+1}, x^i_{t+1}|\pi_t, x^i_t, \gamma^i_t, a^i_t, \psi_t) \\
&= \sum_{\gamma^{-i}_t, x^{-i}_t, a^{-i}_t} \psi_t(\gamma^{-i}_t|\gamma^i_t) 1_{\gamma^{-i}_t(x^{-i}_t)}(a^{-i}_t) \\
&\quad \cdot 1_{\hat{T}(\pi_t, \gamma_t, \boldsymbol{a}_t)}(\pi_{t+1}) Q^i(x^i_{t+1}|x^i_t, \boldsymbol{a}_t),
\end{aligned} \tag{42}$$

which does not depend on $\hat{\phi}^C$. Substituting (41) back to (38b) and expanding (39), one obtains (20). The update equation (21) can be derived by substituting (38b) in (40), utilizing law of iterated expectation as (41), introducing $\psi_t = \hat{\phi}^C_t[\pi_t]$ explicitly to the condition, and dropping $\hat{\phi}^C$ from the superscript due to a reason similar to (42). ∎

F. Proof of Lemma 3

Proof: Due to the limited space, we provide a proof sketch. For an sPCE with a deterministic device $\hat{\phi}^C$, one may construct $\theta_t[\pi_t] := \boldsymbol{\gamma}_t$ for $\hat{\phi}^C_t[\pi_t] = 1_{\boldsymbol{\gamma}_t}(\cdot)$, and verify that (θ, μ) is an sPBE, where μ is the belief system of the given sPCE. It should be straightforward to check the identical belief system and conditional distributions on prescription profiles under this construction. Since any sPBE offers deterministic prescription profiles given a fixed π_t, there is no way for an sPBE to recover the conditional distribution on prescription profiles induced by an sPCE with randomness in $\hat{\phi}^C$. ∎

REFERENCES

[1] H. S. Witsenhausen, "A standard form for sequential stochastic control," *Mathematical Systems Theory*, vol. 7, no. 1, pp. 5–11, 1973.

[2] 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.

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

[4] Y. Ouyang, H. Tavafoghi, and D. Teneketzis, "Dynamic games with asymmetric information: Common information based perfect bayesian equilibria and sequential decomposition," *IEEE Trans. Automatic Control*, vol. 62, no. 1, pp. 222–237, Jan 2017.

[5] D. Vasal, A. Sinha, and A. 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. 81–96, 2019.

[6] D. Tang, H. Tavafoghi, V. Subramanian, A. Nayyar, and D. Teneketzis, "Dynamic games among teams with delayed intra-team information sharing," *Dynamic Games and Applications*, pp. 1–59, 2022.

[7] R. J. Aumann, "Subjectivity and correlation in randomized strategies," *Journal of mathematical Economics*, vol. 1, no. 1, pp. 67–96, 1974.

[8] F. Forges, "Five legitimate definitions of correlated equilibrium in games with incomplete information," *Theory and decision*, vol. 35, pp. 277–310, 1993.

[9] D. Bergemann and S. Morris, "Correlated equilibrium in games with incomplete information," *Cowles Foundation for Research in Economics*, 2011.

[10] F. Forges, "An approach to communication equilibria," *Econometrica*, vol. 54, no. 6, pp. 1375–1385, 1986. [Online]. Available: http://www.jstor.org/stable/1914304

[11] A. Greenwald and M. Zinkevich, "A direct proof of the existence of correlated equilibrium policies in general-sum markov games," *Brown CS: Tech Report CS-05–07*, 2005.

[12] E. Solan and N. Vieille, "Correlated equilibrium in stochastic games," *Games and Economic Behavior*, vol. 38, no. 2, pp. 362–399, 2002.

[13] W. Mao and T. Başar, "Provably efficient reinforcement learning in decentralized general-sum markov games," *Dynamic Games and Applications*, pp. 1–22, 2022.

[14] B. Von Stengel and F. Forges, "Extensive-form correlated equilibrium: Definition and computational complexity," *Mathematics of Operations Research*, vol. 33, no. 4, pp. 1002–1022, 2008.

[15] A. Celli, A. Marchesi, G. Farina, and N. Gatti, "No-regret learning dynamics for extensive-form correlated equilibrium," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 7722–7732. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/5763abe87ed1938799203fb6e8650025-Paper.pdf

[16] P. R. Kumar and P. Varaiya, *Stochastic systems: estimation, identification, and adaptive control*. Englewood Cliffs, NJ: Prentice-Hall, 1986.

[17] H. Tavafoghi, Y. Ouyang, and D. Teneketzis, "A unified approach to dynamic decision problems with asymmetric information: Nonstrategic agents," *IEEE Transactions on Automatic Control*, vol. 67, no. 3, pp. 1105–1119, 2021.

[18] E. Altman, *Constrained Markov decision processes*. Chapman and Hall/CRC, 1999.

[19] H. Guo, X. Liu, H. Wei, and L. Ying, "Online convex optimization with hard constraints: Towards the best of two worlds and beyond," in *Advances in Neural Information Processing Systems*, 2022.