Partially Observable Multi-agent RL with (Quasi-)Efficiency: The Blessing of Information Sharing

Xiangyu Liu [1] **Kaiqing Zhang** [1]

Abstract

We study provable multi-agent reinforcement learning (MARL) in the general framework of partially observable stochastic games (POSGs). To circumvent the known hardness results and the use of computationally intractable oracles, we propose to leverage the potential *information-sharing* among agents, a standard practice in empirical MARL and a common model for multi-agent control systems with communications. We first establish several computation complexity results to justify the necessity of information-sharing, as well as the observability assumption that has enabled quasi-efficient single-agent RL with partial observations, for computational efficiency in solving POSGs. We then propose to further *approximate* the shared common information to construct an approximate model of the POSG, in which planning an approximate equilibrium (in terms of solving the original POSG) can be quasi-efficient, i.e., of quasi-polynomial-time, under the aforementioned assumptions. Furthermore, we develop a partially observable MARL algorithm that is both statistically and computationally quasi-efficient. We hope our study can open up the possibilities of leveraging and even designing different *information structures*, for developing both sample- and computation-efficient partially observable MARL.

1. Introduction

Recent years have witnessed fast development of reinforcement learning (RL) in a wide range of applications, including playing Go games (Silver et al., 2017), robotics (Lillicrap et al., 2016; Long et al., 2018), video games (Vinyals et al., 2019; Berner et al., 2019), and autonomous driving (Shalev-Shwartz et al., 2016; Sallab et al., 2017). Many of these application domains by nature involve *multiple decision-makers* operating in a common environment, with either aligned or misaligned objectives that are affected by their joint behaviors. This has thus inspired surging research interests in multi-agent RL (MARL), with both deeper theoretical and empirical understandings (Busoniu et al., 2008; Zhang et al., 2021a; Hernandez-Leal et al., 2019).

One central challenge of multi-agent learning in these applications is the *imperfection* of information, or more generally, the *partial observability* of environments. Specifically, each agent may possess *different* information about the state and action processes while making decisions. For example, in vision-based multi-robot learning and autonomous driving, each agent only accesses a first-person camera to stream noisy measurements of the object/scene, without accessing the observations or past actions of other agents. This is also referred to as *information asymmetry* in game theory and decentralized decision-making (Nayyar et al., 2013a; Shi et al., 2016). Despite its ubiquity in practice, theoretical understandings of MARL in partially observable settings remain scant. This is somewhat expected since even in single-agent settings, planning and learning under partial observability suffer from well-known computational and statistical hardness results (Papadimitriou & Tsitsiklis, 1987; Mundhenk et al., 2000; Jin et al., 2020). The hardness is known to be amplified for multi-agent decentralized decision-making (Witsenhausen, 1968; Tsitsiklis & Athans, 1985). Existing provable partially observable MARL algorithms either only apply to a small subset of highly structured problems (Zinkevich et al., 2007; Kozuno et al., 2021), or computationally intractable (Liu et al., 2022b).

With these hardness results that can be doubly exponential in the worst case, even a *(quasi-)polynomial* efficient algorithm could represent a non-trivial improvement in partially observable MARL. In particular, we ask and attempt to answer the following question:

Can we have partially observable MARL algorithms that are both *statistically* and *computationally efficient?*

We provide some results towards answering the question positively, by leveraging the potential *information-sharing* among agents, together with a careful compression of the

[1]University of Maryland, College Park. Correspondence to: Kaiqing Zhang <kaiqing@umd.edu>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

shared information. Indeed, the idea of information sharing has been widely used in empirical MARL, e.g., *centralized* training that aggregates all agents' information for more efficient training (Lowe et al., 2017; Rashid et al., 2020); it has also been widely used to model practical multi-agent control systems, e.g., with delayed communications (Witsenhausen, 1971; Nayyar et al., 2010). We defer a thorough literature review to §A, and detail our contributions as follows.

Contributions. We study provable MARL under the general framework of partially observable stochastic games (POSGs), with potential information sharing among agents. First, we establish several computation complexity results of solving POSGs in the presence of *information sharing*, justifying its necessity, together with the necessity of the observability assumption made in the literature. Second, we propose to further approximate the shared common information to construct an *approximate model*, and characterize the computation complexity of planning in this model. We show that for several standard information-sharing structures, a simple *finite-memory* compression can lead to expected approximate common information models in which planning an approximate equilibrium (in terms of solving the original model) has quasi-polynomial time complexity. Third, based on the planning results, we develop a partially observable MARL algorithm that is both statistically and computationally quasi-efficient. To the best of our knowledge, this is the first provably quasi-efficient partially observable MARL algorithm, in terms of both sample and computational complexities. Finally, we also provide experiments to validate: i) the benefit of information sharing as we considered in partially observable MARL; ii) the implementability of our theoretically supported algorithms.

Notation. For two sets B and D, we define $B \setminus D$ as set B minus set D. We use $\emptyset$ to denote the empty set and $[n] := \{1, \cdots, n\}$. For integers $a \le b$, we abbreviate a sequence $(x_a, x_{a+1}, \cdots, x_b)$ by $x_{a:b}$. If $a > b$, then it denotes an empty sequence. When the sequence index starts from m and ends at n, we will treat $x_{a:b}$ as $x_{\max\{a,m\}:\min\{b,n\}}$.

2. Preliminaries

2.1. POSGs and common information

Model. Formally, we define a POSG with n agents by a tuple $\mathcal{G} = (H, \mathcal{S}, \{\mathcal{A}_i\}_{i=1}^n, \{\mathcal{O}_i\}_{i=1}^n, \mathbb{T}, \mathbb{O}, \mu_1, \{r_i\}_{i=1}^n)$, where H denotes the length of each episode, $\mathcal{S}$ is the state space with $|\mathcal{S}| = S$, $\mathcal{A}_i$ denotes the action space for the i^{th} agent with $|\mathcal{A}_i| = A_i$. We denote by $a := (a_1, \cdots, a_n)$ the joint action of all the n agents, and by $\mathcal{A} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_n$ the joint action space with $|\mathcal{A}| = A = \Pi_{i=1}^n A_i$. We use $\mathbb{T} = \{\mathbb{T}_h\}_{h\in[H]}$ to denote the collection of the transition matrices, so that $\mathbb{T}_h(\cdot|s,a) \in \Delta(\mathcal{S})$ gives the probability of the next state if joint action a are taken at state s and step h. In the following discussions, for any given a, we treat $\mathbb{T}_h(a) \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ as a matrix, where each row gives the probability for the next state. We use μ_1 to denote the distribution of the initial state s_1, and $\mathcal{O}_i$ to denote the observation space for the i^{th} agent with $|\mathcal{O}_i| = O_i$. We denote by $o := (o_1, \ldots, o_n)$ the joint observation of all n agents, and by $\mathcal{O} := \mathcal{O}_1 \times \cdots \times \mathcal{O}_n$ with $|\mathcal{O}| = O = \Pi_{i=1}^n O_i$. We use $\mathbb{O} = \{\mathbb{O}_h\}_{h\in[H+1]}$ to denote the collection of the joint emission matrices, so that $\mathbb{O}_h(\cdot|s) \in \Delta(\mathcal{O})$ gives the emission distribution over the joint observation space $\mathcal{O}$ at state s and step h. For notational convenience, we will at times adopt the matrix convention, where $\mathbb{O}_h$ is a matrix with rows $\mathbb{O}_h(\cdot|s_h)$. We also denote $\mathbb{O}_{i,h}(\cdot|s) \in \Delta(\mathcal{O}_i)$ as the marginalized emission for the i^{th} agent. Finally, $r_i = \{r_{i,h}\}_{h\in[H+1]}$ is a collection of reward functions, so that $r_{i,h}(o_h)$ is the reward of the i^{th} agent given the joint observation o_h at step h. This general formulation of POSGs includes several important subclasses. For example, decentralized partially observable Markov decision processes (Dec-POMDPs) are POSGs where the agents share a common reward function, i.e., $r_i = r, \forall i \in [n]$; zero-sum POSGs are POSGs with $n = 2$ and $r_1 = -r_2$. Hereafter, we will use the terminology *cooperative POSG* and *Dec-POMDP* interchangeably.

In a POSG, the states are always hidden from agents, and each agent can only observe its own individual observations. The game proceeds as follows. At the beginning of each episode, the environment samples s_1 from μ_1. At each step h, each agent i observes its own observation $o_{i,h}$, and receives the reward $r_{i,h}(o_h)$ where $o_h := (o_{1,h}, \ldots, o_{n,h})$ is sampled jointly from $\mathbb{O}_h(\cdot|s_h)$. Then each agent i takes the action $a_{i,h}$. After that the environment transitions to the next state $s_{h+1} \sim \mathbb{T}_h(\cdot|s_h, a_h)$. The current episode terminates immediately once s_{H+1} is reached and the reward $r_{i,H+1}(o_{H+1})$ is received. Since the reward at the first step $r_{i,1}(o_{i,1})$ does not depend on the policy, we will assume the trajectory starts from a_1 instead of o_1.

Information sharing, common and private information. Each agent i in the POSG maintains its own information, $\tau_{i,h}$, a collection of historical observations and actions at step h, namely, $\tau_{i,h} \subseteq \{a_1, o_2, \cdots, a_{h-1}, o_h\}$, and the collection of the history at step h is given by $\mathcal{T}_{i,h}$.

In many practical examples (see concrete ones in §3), agents may share part of the history with each other, which may introduce more structure in the game that leads to both sample and computation efficiency. The information sharing splits the history into *common/shared* and *private* information for each agent. The *common information* at step h is a subset of the joint history τ_h: $c_h \subseteq \{a_1, o_2, \cdots, a_{h-1}, o_h\}$, which is available to *all the agents* in the system, and the collection of the common information is denoted as $\mathcal{C}_h$ and define $C_h = |\mathcal{C}_h|$. Given the common information c_h, each agent also has the private information $p_{i,h} = \tau_{i,h} \setminus c_h$, where

the collection of the private information for the i^{th} agent is denoted as $\mathcal{P}_{i,h}$ and its cardinality as $P_{i,h}$. The joint private information at step h is denoted as p_h, where the collection of the joint private history is given by $\mathcal{P}_h = \mathcal{P}_{1,h} \times \cdots \times \mathcal{P}_{n,h}$ and the corresponding cardinality is $P_h = \Pi_{i=1}^n P_{i,h}$. We allow c_h or $p_{i,h}$ to take the special value $\emptyset$ when there is no common or private information. In particular, when $\mathcal{C}_h = \{\emptyset\}$, the problem reduces to the general POSG without any favorable information structure; when $\mathcal{P}_{i,h} = \{\emptyset\}$, every agent holds the same history, and it reduces to a POMDP when the agents share a common reward function and the goal is usually to find the team optimal solution.

Throughout, we also assume that the common information and private information evolve over time properly.

Assumption 1 (Evolution of common and private information)**.** We assume that common information and private information evolve over time as follows:

- Common information c_h is non-decreasing with time, that is, $c_h \subseteq c_{h+1}$ for all h. Let $z_{h+1} = c_{h+1} \setminus c_h$. Thus, $c_{h+1} = \{c_h, z_{h+1}\}$. Further, we have

$$z_{h+1} = \chi_{h+1}(p_h, a_h, o_{h+1}), \tag{2.1}$$

  where χ_{h+1} is a fixed transformation. We use $\mathcal{Z}_{h+1}$ to denote the collection of z_{h+1} at step h. Since we assume the trajectory starts from a_1 instead of o_1, we have $c_1 = \emptyset$.
- Private information evolves according to:

$$p_{i,h+1} = \xi_{i,h+1}(p_{i,h}, a_{i,h}, o_{i,h+1}), \tag{2.2}$$

  where $\xi_{i,h+1}$ is a fixed transformation.

Equation (2.1) states that the increment in the common information depends on the "new" information a_h, o_{h+1} generated between step h and $h+1$ and part of the old information p_h. The increment common information can be implemented by certain sharing and communication protocol among the agents. Equation (2.2) implies that the evolution of private information only depends on the newly generated private information $a_{i,h}$ and $o_{i,h+1}$. These evolution rules are standard in the literature (Nayyar et al., 2013a;b), clearly specifying the source of common information and private information.

2.2. Policies and value functions

We define a stochastic policy for the i^{th} agent at step h as:

$$\pi_{i,h} : \Omega_h \times \mathcal{P}_{i,h} \times \mathcal{C}_h \to \Delta(\mathcal{A}_i). \tag{2.3}$$

The corresponding policy class is denoted as $\Pi_{i,h}$. Hereafter, unless otherwise noted, when referring to *policies*, we mean the policies given in the form of (2.3). Here $\omega_{i,h} \in \Omega_h$ is the random seed, and Ω_h is the random seed space, which is shared among agents. We further denote $\Pi_i = \times_{h\in[H]}\Pi_{i,h}$ as the policy space for agent i and Π as the joint policy space. As a special case, we define the space of deterministic policy as $\widetilde{\Pi}_i$, where $\widetilde{\pi}_i \in \widetilde{\Pi}_i$ maps the private information and common information to an *deterministic* action for the i^{th} agent and the joint space as $\widetilde{\Pi}$.

Another important concept in the common-information-based framework is called the *prescription* (Nayyar et al., 2013b;a), defined for the i^{th} agent as

$$\gamma_{i,h} : \mathcal{P}_{i,h} \to \Delta(\mathcal{A}_i).$$

With such a prescription function, agents can take actions purely based on their local private information. We define $\Gamma_{i,h}$ as the function class for prescriptions, and $\Gamma := \{\Gamma_{i,h}\}_{i\in[n],h\in[H]}$ as the function class for joint prescriptions. Intuitively, the partial function $\pi_{i,h}(\cdot|\,\omega_{i,h}, c_h, \cdot)$ is a prescription given some $\omega_{i,h}$ and c_h. We will define π_i as a sequence of policies for agent i at all steps $h \in [H]$, i.e., $\pi_i = (\pi_{i,1}, \cdots, \pi_{i,H})$ and Π_i as the corresponding collection of policies for agent i. A (potentially correlated) joint policy is denoted as $\pi = \pi_1 \odot \pi_2 \cdots \odot \pi_n \in \Pi$. A product policy is denoted as $\pi = \pi_1 \times \pi_2 \cdots \times \pi_n \in \Pi$ if the distribution of drawing each seed $\omega_{i,h}$ for different agents is independent. For Dec-POMDPs, using stochastic policies will not yield better policies than using only deterministic policies (Oliehoek & Amato, 2016). However, for general POSGs, there might not exist a pure strategy solution in the deterministic policy class. Furthermore, sometimes, we might resort to the general joint policy $\pi = \{\pi_1, \pi_2, \cdots, \pi_h\}$, which could potentially go beyond Π, where π_h is defined as: $\pi_h : \mathcal{A}^{h-1} \times \mathcal{O}^{h-1} \to \Delta(\mathcal{A})$. We denote the collection of such policies as Π^{gen}. For some policy π and event $\mathcal{E}$, we write $\mathbb{P}^{\mathcal{G}}_{s_{1:h}, a_{1:h-1}, o_{2:h}\sim\pi_{1:h-1}}(\mathcal{E})$ to denote the probability of $\mathcal{E}$ when $(s_{1:h}, a_{1:h-1}, o_{2:h})$ is drawn from a trajectory following the policy $\pi_{1:h-1}$ from step 1 to $h-1$ in the model $\mathcal{G}$. We will use the shorthand notation $\mathbb{P}_h^{\pi_{1:h-1},\mathcal{G}}(\cdot)$ if the definition of $(s_{1:h}, a_{1:h-1}, o_{2:h})$ is evident. At times, if the time index h is evident, we will write it as $\mathbb{P}_h^{\pi,\mathcal{G}}(\cdot)$. If the event $\mathcal{E}$ does not depend on the choice of π, we will use $\mathbb{P}_h^{\mathcal{G}}(\cdot)$ and omit π. Similarly, we will write $\mathbb{E}^{\mathcal{G}}_{s_{1:h}, a_{1:h-1}, o_{2:h}\sim\pi}[\cdot]$ to denote expectations and use the shorthand notation $\mathbb{E}^{\mathcal{G}}[\cdot]$ if the expectation does not depend on the choice of π. Furthermore, if we are given some model $\mathcal{M}$ (other than $\mathcal{G}$), the notation of $\mathbb{P}_h^{\mathcal{M}}(\cdot)$, $\mathbb{E}^{\mathcal{M}}[\cdot]$ is defined in the same way. We also denote the indicator of $\mathcal{E}$ as $\mathbb{1}(\mathcal{E}) = 1$ if the event $\mathcal{E}$ is true and 0 otherwise. We will use *strategy* and *policy* interchangeably.

We are now ready to define the *value function* for each agent:

Definition 1 (Value function)**.** For each agent $i \in [n]$ and step $h \in [H]$, given common information c_h and joint π, the value function conditioned on the common information of agent i is defined as: $V_{i,h}^{\pi,\mathcal{G}}(c_h) := \mathbb{E}_\pi^{\mathcal{G}}\left[\sum_{h'=h+1}^{H+1} r_{i,h'}(o_{h'}) \,\middle|\, c_h\right]$, where the expectation is taken over the randomness from the model $\mathcal{G}$, policy π, and the

random seeds. For any $c_{H+1} \in \mathcal{C}_{H+1}$: $V_{i,H+1}^{\pi,\mathcal{G}}(c_{H+1}) := 0$. From now on, we will refer to it as *value function* for short.

Another key concept in our analysis is the belief about the state *and* the private information conditioned on the common information among agents. Formally, at step h, given policies from 1 to $h-1$, we consider the common information-based conditional belief $\mathbb{P}_h^{\pi_{1:h-1},\mathcal{G}}(s_h, p_h | c_h)$. This belief not only infers the current state s_h, but also each agent's private information p_h. With the common-information-based conditional belief, the value function in POSGs has the following recursive structure:

$$V_{i,h}^{\pi,\mathcal{G}}(c_h) = \mathbb{E}_{\pi}^{\mathcal{G}}[r_{i,h+1}(o_{h+1}) + V_{i,h+1}^{\pi,\mathcal{G}}(c_{h+1}) | c_h], \tag{2.4}$$

where the expectation is taken over the randomness of (s_h, p_h, a_h, o_{h+1}) given $\pi_{i,h_{i\in[n]}}$ (and corresponding γ_h). With this, we can define the prescription-value function correspondingly, a generalization of the action-value function in Markov games and MDPs in Definition 11.

2.3. Equilibrium notions

With the definition of the value functions, we can accordingly define the solution concepts. Here we define ϵ-Nash equilibrium (NE) and team-optimal solution as follows, and defer the standard definitions of coarse correlated equilibrium (CCE) and correlated equilibrium (CE) to §B.2.

Definition 2 (ϵ-approximate Nash Equilibrium)**.** For any $\epsilon \geq 0$, a product policy $\pi^\star \in \Pi$ is an ϵ-approximate Nash Equilibrium (NE) of the POSG $\mathcal{G}$ if:

$$\text{NE-gap}(\pi^\star) := \max_i \left(\max_{\pi_i' \in \Pi_i} V_{i,1}^{\pi_i' \times \pi_{-i}^\star, \mathcal{G}}(\emptyset) - V_{i,1}^{\pi^\star,\mathcal{G}}(\emptyset) \right) \leq \epsilon.$$

Definition 3 (ϵ-approximate team-optimal policy in Dec-POMDPs with information-sharing structures)**.** When the reward functions $r_{i,h}$ are identical for all $i \in [n]$, i.e., $r_{i,h} = r_h$, and the POSG reduces to a Dec-POMDP, then a policy $\pi^\star \in \widetilde{\Pi}$ is a team optimal policy if: $V_1^{\pi^\star,\mathcal{G}}(\emptyset) \geq \max_{\pi' \in \widetilde{\Pi}} V_1^{\pi',\mathcal{G}}(\emptyset) - \epsilon$.

By restricting to deterministic policies, it does not lose any optimality (Nayyar et al., 2013b). It is also worth noting that, the team-optimal solution is always a NE, a NE is always a CE, and a CE is always a CCE.

3. Information Sharing in Applications

The aforementioned information-sharing structure can indeed be common in real-world applications. For example, for a self-driving car to avoid collision and successfully navigate, the other cars from the same fleet/company would necessarily communicate with each other (possibly with delays) about the road situation. The separation between common information and private information then arises naturally (Gong et al., 2016). Similar examples can also be found in cloud computing and power systems (Altman et al., 2009). Here, we outline several representative information-sharing structures that fit into our algorithmic framework, and defer more examples in §B.4.

Example 1 (One-step delayed sharing)**.** At any step $h \in [H+1]$, the common and private information are given as $c_h = \{o_{2:h-1}, a_{1:h-1}\}$ and $p_{i,h} = \{o_{i,h}\}$, respectively. In other words, the players share all the action-observation history until the previous step $h-1$, with only the new observation being private information. This model has been shown useful for power control (Altman et al., 2009).

Example 2 (State controlled by one controller with asymmetric delay sharing)**.** We assume there are 2 players for convenience. It extends naturally to n-player settings. Consider the case where the state dynamics are controlled by player 1, i.e., $\mathbb{T}_h(\cdot | s_h, a_{1,h}, a_{2,h}) = \mathbb{T}_h(\cdot | s_h, a_{1,h}, a'_{2,h})$ for all $s_h, a_{1,h}, a_{2,h}, a'_{2,h}, h$. There are two kinds of delay-sharing structures we could consider: **Case A:** the information structure is given as $c_h = \{o_{1,2:h}, o_{2,2:h-d}, a_{1,1:h-1}\}$, $p_{1,h} = \emptyset$, $p_{2,h} = \{o_{2,h-d+1:h}\}$, i.e., player 1's observations are available to player 2 instantly, while player 2's observations are available to player 1 with a delay of $d \geq 1$ time steps. **Case B:** similar to **Case A** but player 1's observation is available to player 2 with a delay of 1 step. The information structure is given as $c_h = \{o_{1,2:h-1}, o_{2,2:h-d}, a_{1,1:h-1}\}$, $p_{1,h} = \{o_{1,h}\}$, $p_{2,h} = \{o_{2,h-d+1:h}\}$, where $d \geq 1$. This kind of asymmetric sharing is common in network routing (Pathak et al., 2008), where packages arrive at different hosts with different delays, leading to asymmetric delay sharing among hosts.

Example 3 (Symmetric information game)**.** Consider the case when all observations and actions are available for all the agents, and there is no private information. Essentially, we have $c_h = \{o_{2:h}, a_{1:h-1}\}$ and $p_{i,h} = \emptyset$. We will also denote this as *fully sharing* hereafter.

4. Hardness and Planning with Exact Model

4.1. Hardness on finding team-optimum and equilibria

Throughout, we mainly consider the NE, CE, and CCE as our solution concepts. However, in Dec-POMDPs, a special class of POSGs with common rewards, a more common and favorable objective is the *team optimum*. However, in Dec-POMDPs, it is known that computing even approximate team optimal policies is `NEXP-complete` (Bernstein et al., 2002; Rabinovich et al., 2003), i.e., the algorithms as in (Hansen et al., 2004) may take doubly exponential time in the worst case. Such hardness cannot be circumvented under our information-sharing framework either without further assumptions, since even with fully

sharing, the problem becomes solving a POMDP to its optimum, which is still PSPACE-complete (Papadimitriou & Tsitsiklis, 1987).

Recently, (Golowich et al., 2022b) considers *observable* POMDPs that rule out the ones with uninformative observations, for which computationally (quasi)-efficient algorithms can be developed. In the hope of obtaining computational (quasi)-efficiency for POSGs (including Dec-POMDPs), we could make a similar observability assumption on the joint observations as below. Note that this observability assumption is equivalent (up to a factor of at most $\sqrt{O}$) to the ϵ-weakly revealing condition in (Liu et al., 2022b), under which there also exists a *statistically* efficient algorithm for solving POSGs.

Assumption 2 (γ-Observability)**.** Let $\gamma > 0$. For $h \in [H]$, we say that the matrix $\mathbb{O}_h$ satisfies the γ-observability assumption if for each $h \in [H]$, for any $b, b' \in \Delta(\mathcal{S})$,

$$\left\|\mathbb{O}_h^\top b - \mathbb{O}_h^\top b'\right\|_1 \geq \gamma \left\|b - b'\right\|_1.$$

A POSG (Dec-POMDP) satisfies (one-step) γ-observability if all its $\mathbb{O}_h$ for $h \in [H]$ do so.

However, we show that even under such an assumption, and with the favorable 1-step delayed sharing structure as introduced in §3, computing team optimal policy in Dec-POMDPs can be NP-Hard. Moreover, we show that missing either Assumption 2 or any information-sharing structures will make the problem of even computing NE/CE/CCE, the more relaxed solution concepts than team optimum in Dec-POMDPs, PSAPCE-Hard. This shows the necessity of both Assumption 2 and certain information-sharing structures. Next, we will focus on planning and learning in POSGs under these assumptions.

4.2. Planning with strategy-independent common belief

For optimal/equilibrium policy computation, it is known that backward induction is one of the most useful approaches for solving (fully-observable) Markov games. However, the essential impediment to applying backward induction in *asymmetric* information games is the fact that a player's posterior beliefs about the system state and about other players' information may depend on the *strategies* used by the players in the past. If the nature of system dynamics and the information structure of the game ensure that the players' posterior beliefs are *strategy independent*, then a backward induction can be derived for equilibrium computation (Nayyar et al., 2013a; Gupta et al., 2014). We formalize this conceptual argument as the following assumption.

Assumption 3 (Strategy independence of beliefs)**.** Consider any step $h \in [H]$, any choice of joint policies $\pi \in \Pi$, and any realization of common information c_h that has a nonzero probability under the trajectories generated by $\pi_{1:h-1}$. Consider any other policies $\pi'_{1:h-1}$, which also give a nonzero probability to c_h. Then, we assume that: for any such $c_h \in \mathcal{C}_h$, and any $p_h \in \mathcal{P}_h, s_h \in \mathcal{S}$, $\mathbb{P}_h^{\pi_{1:h-1},\mathcal{G}}(s_h, p_h | c_h) = \mathbb{P}_h^{\pi'_{1:h-1},\mathcal{G}}(s_h, p_h | c_h)$.

This assumption has been made in the literature (Nayyar et al., 2013a; Gupta et al., 2014), which is related to the notion of *one-way separation* in stochastic control, that is, the estimation (of the state in standard stochastic control and of the state and private information) in Assumption 3 is *independent* of the control strategy. A naive attempt to relaxing this is to also include the past $\pi_{1:h-1}$ in addition to c_h when computing the belief of states and private information. In other words, one can firstly find a solution $\pi^\star = \{\pi_h^\star\}_{h\in[H]}$, where the execution of $\pi_h^\star$ depends on the past $\pi_{1:h-1}^\star$. Then one can eliminate such dependency through a methods called "forward-sweeping" to find some policy $\widehat{\pi}^\star$ so that $V_i^{\pi^\star,\mathcal{G}}(\emptyset) = V_i^{\widehat{\pi}^\star,\mathcal{G}}(\emptyset)$, and $\widehat{\pi}^\star$ can be executed in a decentralized way. In fact, such an idea turns out to be useful for computing team optimal policies in Dec-POMDPs (Nayyar et al., 2013b), but not effective for finding equilibrium in the game setting, since one joint policy's value being equal to that at an equilibrium does not necessarily imply it is also an equilibrium policy. For more detailed discussion, we refer to (Nayyar et al., 2013a). There are also works not requiring this assumption (Ouyang et al., 2016; Tavafoghi et al., 2016), but under a different perfect Bayesian equilibrium framework. We leave the study of developing computationally tractable approaches under this framework as our future work. Before proceeding with further analysis, It is worth mentioning that examples introduced in §3 all satisfy this assumption (see (Nayyar et al., 2013a) and also §E.4).

With Assumption 3, we are able to develop a planning algorithm (shown in Algorithm 1) with the following time complexity. The algorithm is based on value iteration on the common information space, which runs in a backward way, enumerating all possible c_h at each step h and computing the corresponding equilibrium in the prescription space. Detailed description of the algorithm is deferred to §C.

Theorem 1. Fix $\epsilon > 0$. For the POSG $\mathcal{G}$ with information structure satisfying Assumption 3, given access to the belief $\mathbb{P}_h^{\mathcal{G}}(s_h, p_h | c_h)$, Algorithm 1 computes an ϵ-NE if $\mathcal{G}$ is zero-sum or cooperative, and an ϵ-CE/CCE if $\mathcal{G}$ is general-sum, with time complexity $\max_{h\in[H]} C_h \cdot \mathrm{poly}(S, A, P_h, H, \frac{1}{\epsilon})$.

This theorem characterizes the dependence of computation complexity on the cardinality of the common information set and private information set. Ideally, to get a computationally efficient algorithm, we should *design* the information-sharing strategy such that C_h and P_h are both small. To get a sense of how large $C_h P_h$ could be, we consider one common

scenario where each player has *perfect recall*, i.e., she remembers what she did in prior moves, and also remembers everything that she knew before.

Definition 4 (Perfect recall)**.** We say that player i has perfect recall if for any $h \in [H+1]$, it holds that $\{a_{i,1:h-1}, o_{i,2:h}\} \subseteq \tau_{i,h}$, and $\tau_{i,h} \subseteq \tau_{i,h+1}$.

If each player has perfect recall as defined above, we can show that $C_h P_h$ must be exponential in the horizon index h.

Lemma 1. Fix any $h \in [H]$. If each player has perfect recall as given in Definition 4, then for any information-sharing structures $C_h P_h \geq (OA)^{h-1}$.

With this result, we know that the computation complexity of our planning algorithm must suffer from the exponential dependence of $\Omega((OA)^h)$. This negative result tells us it is barely possible to get computational efficiency for running planning algorithms in the true model $\mathcal{G}$, since the $C_h P_h$ has to be very large oftentimes. It is worth noting that for obtaining this result (Theorem 1), we did not utilize our Assumption 2. Thus, this negative result is consistent with our hardness results in Proposition 3.

5. Planning and Learning with Approximate Common Information

5.1. Computationally (quasi-)efficient planning

Previous exponential complexity comes from the fact that C_h and P_h could not be made simultaneously small under the common scenario with perfect recall. To address this issue, we propose to further *compress* the information available to the agent with certain regularity conditions, while approximately maintaining the optimality of the policies computed/learned from the compressed information. Notably, there is a tradeoff between *compression error* and *computational tractability*. We will show next that by properly compressing only the common information, we can obtain efficient planning (and learning) algorithms with favorable suboptimality guarantees. To introduce the idea more formally, we first define the concept of *approximate* common information model in our settings.

Definition 5 (Approximate common information)**.** We define an *expected approximate common information model* of $\mathcal{G}$ as

$$\mathcal{M} := \left(\{\widehat{\mathcal{C}}_h\}_{h\in[H+1]}, \{\widehat{\phi}_{h+1}\}_{h\in[H]}, \{\mathbb{P}_h^{\mathcal{M},z}, \mathbb{P}_h^{\mathcal{M},o}\}_{h\in[H]}, \Gamma, \widehat{r} \right),$$

where Γ is the function class for joint prescriptions, $\mathbb{P}_h^{\mathcal{M},z} : \widehat{\mathcal{C}}_h \times \Gamma_h \to \Delta(\mathcal{Z}_{h+1})$, gives the probability of z_{h+1} under given $\widehat{c}_h \in \widehat{\mathcal{C}}_h$, where $\mathcal{Z}_{h+1}$ is the space of increment common information, and $\{\gamma_{i,h}\}_{i\in[n]} \in \Gamma_h$. Similarly, $\mathbb{P}_h^{\mathcal{M},o} : \widehat{\mathcal{C}}_h \times \Gamma_h \to \Delta(\mathcal{O})$ gives the probability of o_{h+1} under given $\widehat{c}_h \in \widehat{\mathcal{C}}_h$, and $\{\gamma_{i,h}\}_{i\in[n]} \in \Gamma_h$. We denote $\widehat{C}_h := |\widehat{\mathcal{C}}_h|$ for any $h \in [H+1]$. We say $\mathcal{M}$ is an $(\epsilon_r(\mathcal{M}), \epsilon_z(\mathcal{M}))$-*expected-approximate common information model* of $\mathcal{G}$ with the *approximate common information* defined by $\{\widehat{c}_h\}_{h\in[H]}$ for some compression function $\widehat{c}_h = \text{Compress}_h(c_h)$, if it satisfies the following:

- It evolves in a recursive manner, i.e., for each h there exists a transformation function $\widehat{\phi}_{h+1}$ such that

$$\widehat{c}_{h+1} = \widehat{\phi}_{h+1}(\widehat{c}_h, z_{h+1}), \tag{5.1}$$

  where we recall that $z_{h+1} = c_{h+1} \setminus c_h$ is the new common information added at step h.
- It suffices for approximate performance evaluation, i.e., for any prescription γ_h and joint policy $\pi' \in \Pi^{\text{gen}}$

$$\mathbb{E}^{\mathcal{G}}_{a_{1:h-1}, o_{2:h}\sim\pi'} \Big| \mathbb{E}^{\mathcal{G}}[r_{i,h+1}(o_{h+1}) \mid c_h, \gamma_h] - \mathbb{E}^{\mathcal{M}}[\widehat{r}_{i,h+1}(o_{h+1}) \mid \widehat{c}_h, \gamma_h] \Big| \leq \epsilon_r(\mathcal{M}). \tag{5.2}$$

- It suffices for approximately predicting the common information, i.e., for any prescription γ_h and joint policy $\pi' \in \Pi^{\text{gen}}$, and for $\mathbb{P}_h^{\mathcal{G}}(z_{h+1} \mid c_h, \gamma_h)$ and $\mathbb{P}_h^{\mathcal{M},z}(z_{h+1} \mid \widehat{c}_h, \gamma_h)$, we have

$$\mathbb{E}^{\mathcal{G}}_{a_{1:h-1}, o_{2:h}\sim\pi'} \left\| \mathbb{P}_h^{\mathcal{G}}(\cdot \mid c_h, \gamma_h) - \mathbb{P}_h^{\mathcal{M},z}(\cdot \mid \widehat{c}_h, \gamma_h) \right\|_1 \leq \epsilon_z(\mathcal{M}). \tag{5.3}$$

Remark 1. $\mathcal{M}$ defined above is indeed a *Markov game*, where the state space is $\{\widehat{\mathcal{C}}_h\}_{h\in[H+1]}$, Γ is the joint action space, $\{\mathbb{P}_h^{\mathcal{M},z}\}_{h\in[H]}$ together with $\{\widehat{\phi}_{h+1}\}_{h\in[H]}$ is the transition, and $\mathbb{E}^{\mathcal{M}}[\widehat{r}_{i,h+1}(o_{h+1}) \mid \widehat{c}_h, \gamma_h]$ is the reward given *state* $\widehat{c}_h$ and *joint action* γ_h.

Remark 2. Note that our requirement of approximate information in the definition can be much weaker than the existing and related definition (Kao & Subramanian, 2022; Mao et al., 2020; Subramanian et al., 2022), which requires the *total variation distance* between $\mathbb{P}_h^{\mathcal{G}}(\cdot \mid c_h, \gamma_h)$ and $\mathbb{P}_h^{\mathcal{M},z}(\cdot \mid \widehat{c}_h, \gamma_h)$ to be *uniformly* bounded for *all* c_h. In contrast, we only require such distance to be small *in expectation*. In fact, the kind of compression in (Kao & Subramanian, 2022; Mao et al., 2020; Subramanian et al., 2022) may be unnecessary and computationally intractable when it comes to efficient planning. Firstly, some common information may have very low visitation frequency under *any* policy π, which means that we can allow large variation between true common belief and approximate common belief for these c_h, which are inherently less important for the decision-making problem. Secondly, even in the single-agent setting, where $c_h = \{a_{1:h-1}, o_{2:h}\}$, the size of such approximate information with errors uniformly bounded

for all $\{a_{1:h-1}, o_{2:h}\}$ could not be sub-exponential, as shown by Example B.2 in (Golowich et al., 2022b). Therefore, for some kinds of common information, it is actually not possible to reduce the order of complexity through the approximate common belief with errors uniformly bounded. Requiring only expected approximation errors to be small is one key to enabling our efficient planning approach next.

Although we have characterized what conditions the expected approximate common information model $\mathcal{M}$ should satisfy to well approximate the underlying $\mathcal{G}$, it is, in general, unclear how to *construct* such an $\mathcal{M}$, mainly how to define $\{\mathbb{P}_h^{\mathcal{M},z}, \mathbb{P}_h^{\mathcal{M},o}\}_{h\in[H]}$, even if we already know how to compress the common information. To address this, in the following definition, we provide a way to construct $\{\mathbb{P}_h^{\mathcal{M},z}, \mathbb{P}_h^{\mathcal{M},o}\}_{h\in[H]}$ by an approximate belief over states and private information $\{\mathbb{P}_h^{\mathcal{M},c}(s_h, p_h|\widehat{c}_h)\}_{h\in[H]}$ by Definition 13, which will facilitate the construction for $\mathcal{M}$ later.

It is direct to verify that we can construct an expected approximate common information model $\mathcal{M}(\mathcal{G})$ for $\mathcal{G}$ such that $\epsilon_z(\mathcal{M}(\mathcal{G})) = \epsilon_r(\mathcal{M}(\mathcal{G})) = 0$, where in this $\mathcal{M}(\mathcal{G})$, we have $\widehat{c}_h = c_h$ for any $h \in [H+1], c_h \in \mathcal{C}_h, \widehat{r} = r$, and $\mathcal{M}(\mathcal{G})$ is consistent with $\{\mathbb{P}_h^{\mathcal{G}}(s_h, p_h|c_h)\}_{h\in[H]}$. Without ambiguity, we will use the shorthand notation ϵ_r, ϵ_z for $\epsilon_r(\mathcal{M}), \epsilon_z(\mathcal{M})$, respectively. With such an expected approximate common information model, similar to Algorithm 1, we develop a value-iteration-type algorithm (Algorithm 3) running on the model $\mathcal{M}$ instead of $\mathcal{G}$, which outputs an approximate NE/CE/CCE, enjoying the following guarantees.

Theorem 2. Fix $\epsilon_r, \epsilon_z, \epsilon_e > 0$. Suppose there exists an (ϵ_r, ϵ_z)-expected-approximate common information model $\mathcal{M}$ for the POSG $\mathcal{G}$ that satisfies Assumption 3. Furthermore, if $\mathcal{M}$ is consistent with some given approximate belief $\{\mathbb{P}_h^{\mathcal{M},c}(s_h, p_h|\widehat{c}_h)\}_{h\in[H]}$, then Algorithm 3 outputs a $\widehat{\pi}^\star$ such that NE-gap$(\widehat{\pi}^\star) \le 2H\epsilon_r + H^2\epsilon_z + H\epsilon_e$ if $\mathcal{G}$ is zero-sum or cooperative, and CE/CCE-gap$(\widehat{\pi}^\star) \le 2H\epsilon_r + H^2\epsilon_z + H\epsilon_e$ if $\mathcal{G}$ is general-sum, where the time complexity is $\max_{h\in[H]} \widehat{C}_h \cdot \mathrm{poly}(S, A, P_h, H, \frac{1}{\epsilon_e})$.

The detailed description of the algorithm is deferred to §C and the consistency between the model and belief is defined in Definition 13. As a sanity check, it is easy to see that if we use previous $\mathcal{M}(\mathcal{G})$ as the expected approximate common information model with the *uncompressed* common information such that $\epsilon_z(\mathcal{M}(\mathcal{G})) = \epsilon_r(\mathcal{M}(\mathcal{G})) = 0$, then by letting $\epsilon_e = \frac{\epsilon}{H}$, we recover our Theorem 1. Then Theorem 2 shows that one could use a compressed version, if it exists, instead of the exact common information, to compute approximate NE/CE/CCE, with the quantitative characterization of the error bound due to this compression. To get an overview of our algorithmic framework, we also refer to Figure 2.

Criteria of information-sharing design for efficient planning. Now we sketch the criterion of designing the information-sharing strategy for efficient planning:

- $\{\mathcal{C}_h\}_{h\in[H+1]}$ satisfies Assumption 3.
- Cardinality of $\{\mathcal{P}_h\}_{h\in[H+1]}$ should be small.
- Cardinality of $\{\widehat{\mathcal{C}}_h\}_{h\in[H+1]}$ should be small.
- Construction of the expected approximate common information model $\mathcal{M}$, i.e., the construction of $\mathbb{E}^{\mathcal{M}}[\widehat{r}_{i,h+1}(o_{h+1})|\widehat{c}_h, \gamma_h]$ and $\mathbb{P}_h^{\mathcal{M},z}(\cdot|\widehat{c}_h, \gamma_h)$ should be computationally efficient.

Planning in observable POSGs without intractable oracles. Theorem 2 applies to any expected approximate common information model as given in Definition 5, by substituting the corresponding $\widehat{C}_h$. Note that it does not provide a way to *construct* such expected approximate common information models that ensure the computation complexity in the theorem is (quasi-)polynomial.

Next, we show that in several natural and standard information structure examples, a simple *finite-memory* compression can attain the goal of computing ϵ-NE/CE/CCE without intractable oracles, where we refer to §E.4 for the concrete form of the finite memory compression. Based on this, we present the corresponding quasi-polynomial time complexities as follows.

Theorem 3. Fix $\epsilon > 0$. Under Assumption 2, there exists a quasi-polynomial time algorithm that can compute ϵ-NE if $\mathcal{G}$ is zero-sum or cooperative, and an ϵ-CE/CCE if $\mathcal{G}$ is general-sum, with the information-sharing structures in §3.

5.2. Statistically (quasi-)efficient learning

Until now, we have been assuming the full knowledge of the model $\mathcal{G}$ (the transition kernel, observation emission, and reward functions). In this full-information setting, we are able to construct some model $\mathcal{M}$ to approximate the true $\mathcal{G}$ according to the conditions we identified in Definition 5. However, when we only have access to the samples drawn from the POSG $\mathcal{G}$, it is difficult to directly construct such a model due to the lack of the model specification. To address this issue, the solution is to construct a specific expected approximate common information model that depends on the policies that generate the data for such a construction, which can be denoted by $\widetilde{\mathcal{M}}(\pi^{1:H})$. For such a model, one could simulate and sample by running policies $\pi^{1:H}$ in the true model $\mathcal{G}$. To introduce such a model $\widetilde{\mathcal{M}}(\pi^{1:H})$, we have the following formal definition.

Definition 6 (Policy-dependent expected approximate common information model)**.** Given H joint policies $\pi^{1:H}$, where $\pi^h \in \Pi^{\mathrm{gen}}$ for $h \in [H]$ and approximate reward functions $\widehat{r}$, we define the policy-dependent approximate common information model $\widetilde{\mathcal{M}}(\pi^{1:H})$ such that it is consistent

with the *policy-dependent* belief $\{\mathbb{P}_h^{\pi^h,\mathcal{G}}(s_h, p_h|\widehat{c}_h)\}_{h\in[H]}$ as per Definition 13.

The key to the definition above resorts to an approximate common information-based conditional belief $\{\mathbb{P}_h^{\pi^h,\mathcal{G}}(s_h, p_h|\widehat{c}_h)\}_{h\in[H]}$ that is defined by running policy $\pi^h \in \Pi^{\text{gen}}$ in $\mathcal{G}$. In particular, we have the following fact.

Proposition 1. Given $\widetilde{\mathcal{M}}(\pi^{1:H})$ as in Definition 6, it holds that for any $h \in [H]$, $\widehat{c}_h \in \widehat{\mathcal{C}}_h$, $\gamma_h \in \Gamma_h$, $o_{h+1} \in \mathcal{O}$, $z_{h+1} \in \mathcal{Z}_{h+1}$: $\mathbb{P}_h^{\widetilde{\mathcal{M}}(\pi^{1:H}),z}(z_{h+1}|\widehat{c}_h, \gamma_h) = \mathbb{P}_h^{\pi^h,\mathcal{G}}(z_{h+1}|\widehat{c}_h, \gamma_h)$, $\mathbb{P}_h^{\widetilde{\mathcal{M}}(\pi^{1:H}),o}(o_{h+1}|\widehat{c}_h, \gamma_h) = \mathbb{P}_h^{\pi^h,\mathcal{G}}(o_{h+1}|\widehat{c}_h, \gamma_h)$.

This proposition verifies that we can have access to the *samples* from the transition and reward of $\widetilde{\mathcal{M}}(\pi^{1:H})$ at step h, by executing the policy π^h in the underlying model $\mathcal{G}$. Now we are ready to present the main theorem for learning such an expected approximate common information model $\widetilde{\mathcal{M}}(\pi^{1:H})$. A major difference from the analysis for planning is that in the learning setting, we need to *explore* the space of approximate common information, which is the function of a sequence of observations and actions, and we need to characterize the *length* of the approximate common information as defined below.

Definition 7 (Length of approximate common information)**.** Given $\{\widehat{\mathcal{C}}_h\}_{h\in[H+1]}$, define the integer $\widehat{L} \geq 0$ as the minimum length such that for each $h \in [H+1]$ and each $\widehat{c}_h \in \widehat{\mathcal{C}}_h$, there exists some mapping $\widehat{f}_h : \mathcal{A}^{\widehat{L}} \times \mathcal{O}^{\widehat{L}} \to \widehat{\mathcal{C}}_h$ and the sequence $x_h = \{a_{\max\{h-\widehat{L},1\}}, o_{\max\{h-\widehat{L},1\}+1}, \cdots, o_h\}$, such that $\widehat{f}_h(x_h) = \widehat{c}_h$.

Such an $\widehat{L}$ characterizes the length of the constructed approximate common information, for which our final sample complexity would necessarily depend on, since we need to do explorations for the steps after $h - \widehat{L}$. It is worth noting that such an $\widehat{L}$ always exists since we can always set $\widehat{L} = H$, and there always exists the mapping $\widehat{f}_h$ such that $\widehat{f}_h(a_{1:h-1}, o_{2:h}) = \widehat{c}_h$, where $\widehat{f}_h$ is a composition of the mapping from $\{a_{1:h-1}, o_{2:h}\}$ to c_h, which is given by the evolution rules in Definition 1, and the compression function Compress_h, the mapping from c_h to $\widehat{c}_h$. With this definition of $\widehat{L}$, we propose Algorithm 5, which learns the model $\widetilde{\mathcal{M}}(\pi^{1:H})$, mainly the two quantities in Proposition 1 by executing policies $\pi^{1:H}$ in the true model $\mathcal{G}$ with the following sample complexity depending on $\widehat{L}$.

Theorem 4. Suppose the POSG $\mathcal{G}$ satisfies Assumption 3. Given compression function of common information, $\text{Compress}_h : \mathcal{C}_h \to \widehat{\mathcal{C}}_h$ for $h \in [H]$, $\widehat{L}$ is as defined in Definition 7. For any H policies $\pi^{1:H}$, where $\pi^h \in \Pi^{\text{gen}}$, $\pi^h_{h-\widehat{L}:h} = \text{Unif}(\mathcal{A})$ for $h \in [H]$, and approximate reward functions $\widehat{r} = \{\{\widehat{r}_{i,h}\}_{i=1}^n\}_{h=1}^H$, we assume $\widetilde{\mathcal{M}}(\pi^{1:H})$ is an $(\epsilon_r(\pi^{1:H}, \widehat{r}), \epsilon_z(\pi^{1:H}))$ expected approximate common information model of $\mathcal{G}$. Fix some parameters $\delta_1, \theta_1, \theta_2, \zeta_1, \zeta_2 > 0$ for Algorithm 5, $\epsilon_e > 0$ for Algorithm 3, and $\phi > 0$, define the approximation error for estimating $\widetilde{\mathcal{M}}(\pi^{1:H})$ using samples under the policy $\pi^{1:H}$ as $\epsilon_{apx}(\pi^{1:H}, \widehat{L}, \zeta_1, \zeta_2, \theta_1, \theta_2, \phi)$. Then Algorithm 5, together with Algorithm 3, can learn $\widetilde{\mathcal{M}}(\pi^{1:H})$ with the sample complexity $N_0 = \text{poly}(\max_h P_h, \max_h \widehat{C}_h, H, A, O, \frac{1}{\zeta_1}, \frac{1}{\zeta_2}, \frac{1}{\theta_1}, \frac{1}{\theta_2}) \log \frac{1}{\delta_1}$, and output an ϵ-NE if $\mathcal{G}$ is zero-sum or cooperative, and an ϵ-CE/CCE if $\mathcal{G}$ is general-sum, with probability at least $1 - \delta_1$, where $\epsilon := H\epsilon_r(\pi^{1:H}, \widehat{r}) + H^2\epsilon_z(\pi^{1:H}) + (H^2 + H)\epsilon_{apx}(\pi^{1:H}, \widehat{L}, \zeta_1, \zeta_2, \theta_1, \theta_2, \phi) + H\epsilon_e$.

A detailed version of this theorem is deferred to §D.3. This meta-theorem presents a sample complexity guarantee of learning expected approximate common information model $\widetilde{\mathcal{M}}(\pi^{1:H})$ under the approximate common-information framework, in the model-free setting. It is agnostic to the choice of approximate common information $\widehat{c}_h$, policies $\pi^{1:H}$, and approximate reward function $\widehat{r}$. Therefore, the final results will necessarily depend on the three error terms $\epsilon_r(\pi^{1:H}, \widehat{r})$, $\epsilon_z(\pi^{1:H})$, and $\epsilon_{apx}(\pi^{1:H}, \widehat{L}, \zeta_1, \zeta_2, \theta_1, \theta_2, \phi)$, which will be instantiated for different examples later to obtain both sample and time complexity results. As before, the meta-theorem applies to cases beyond the following examples, as long as one can compress the common information properly. The following examples just happen to be the ones where a simple finite-memory truncation can give us desired complexities.

Sample (quasi-)efficient learning in POSGs without intractable oracles. Now we apply the meta-theorem, Theorem 4, and obtain polynomial sample and quasi-polynomial time complexities for learning the ϵ-NE/CE/CCE, under several standard information structures.

Theorem 5. Fix $\epsilon > 0$. Under Assumption 2, there exists a multi-agent RL algorithm that learns an ϵ-NE if $\mathcal{G}$ is zero-sum or cooperative, and an ϵ-CE/CCE if $\mathcal{G}$ is general-sum, with information-sharing structures in §3, in time and sample complexities that are *both* quasi-polynomial.

A full version of the theorem is presented in §D.3, with proof provided in §E.5. Note that our algorithm is computationally (quasi-)efficient, in contrast to the only existing sample-efficient MARL algorithm for POSGs in (Liu et al., 2022b), which uses computationally intractable oracles.

Algorithm description. We briefly introduce the algorithm that achieves the guarantees in Theorem 5, i.e., Algorithm 7, and defer more details to §C due to space limitation. The first step is to find the approximate reward function $\widehat{r}$ and policies $\pi^{1:H}$ such that the three error terms $\epsilon_{apx}(\pi^{1:H}, \widehat{L}, \zeta_1, \zeta_2, \theta_1, \theta_2, \phi)$, $\epsilon_r(\pi^{1:H}, \widehat{r})$, and $\epsilon_z(\pi^{1:H})$ in

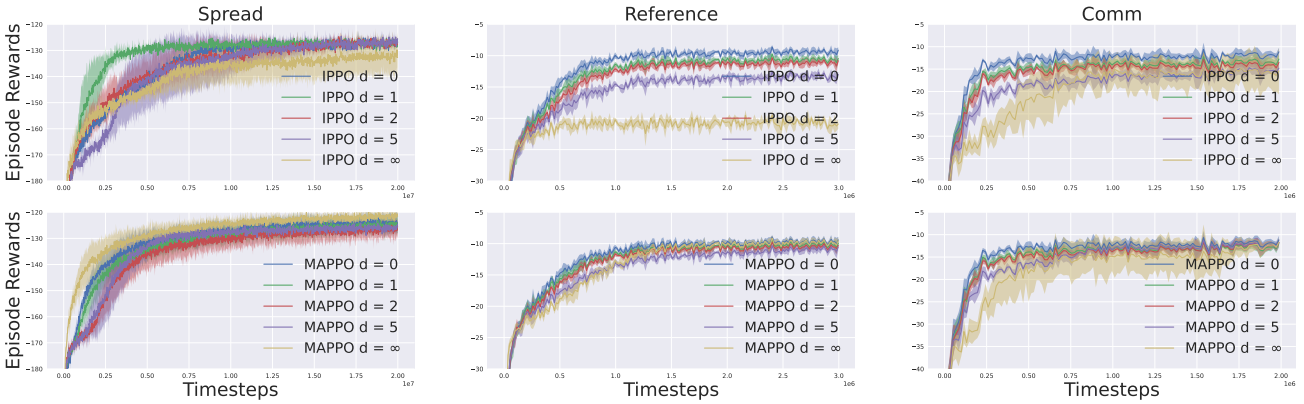

Figure 1. Performance of MAPPO and IPPO in various delayed-sharing settings.

Theorem 4 are minimized. It turns out that the three errors can be minimized using Barycentric-Spanner-based (Awerbuch & Kleinberg, 2008; Golowich et al., 2022a) exploration techniques. The next step is to learn the empirical estimate $\widehat{\mathcal{M}}(\pi^{1:H})$ of $\widetilde{\mathcal{M}}(\pi^{1:H})$, by exploring the approximate common information space $\{\widehat{\mathcal{C}}_h\}_{h\in[H+1]}$ using Algorithm 5. The key to exploring the approximate common information space is to take uniformly random actions from step $h-\widehat{L}$ to h, which has been used for exploration in finite-memory based state spaces in existing works (Uehara et al., 2022; Efroni et al., 2022; Golowich et al., 2022a). Once such a $\widehat{\mathcal{M}}(\pi^{1:H})$ is constructed, we run our planning algorithms (developed in §5.1) on $\widehat{\mathcal{M}}(\pi^{1:H})$ to compute an approximate NE/CE/CCE. The final step is to do policy evaluation to select the equilibrium policy to output, since for the first step we may only obtain a set of $\{\pi^{1:H,j},\widehat{r}^j\}_{j\in[K]}$ for some integer $K>0$ and only some of them can minimize the three error terms. Specifically, for any given policy $\pi^\star$ and $i\in[n]$, the key idea of policy evaluation is that we compute its best response introduced in Algorithm 4 in all the models $\{\widehat{\mathcal{M}}(\pi^{1:H,j})\}_{j\in[K]}$, where $K=\mathrm{poly}(H,S)$ to get $\{\pi^{\star,j}_{-i}\}_{j\in[K]}$ and select the one $\pi^{\star,\widehat{j}}_{-i}$ for some $\widehat{j}$ with the highest empirical rewards by rolling them out in the true $\mathcal{G}$. With the guarantee that there must be a $j\in[K]$ such that $\widehat{\mathcal{M}}(\pi^{1:H,j})$ is a good approximation of $\mathcal{G}$ in the sense of Definition 5, it can be shown that $\pi^{\star,\widehat{j}}_{-i}$ will be an approximate best response in $\mathcal{G}$ with high probability. With the best-response policy, we can select the equilibrium policy with the lowest NE/CE/CCE-gap, which turns out to be an approximate NE/CE/CCE in $\mathcal{G}$.

6. Experimental Results

Information sharing improves performance. We consider the popular deep MARL benchmarks, multi-agent particle-world environment (MPE) (Lowe et al., 2017). We train both state-of-the-art centralized-training algorithm MAPPO and decentralized-training algorithm IPPO (Yu et al., 2021) with different information-sharing mechanisms by varying the information-sharing delay from 0 to ∞. Note that the original algorithm in (Yu et al., 2021) corresponds to the case, where the delay is $d=\infty$. The rewards during training are shown in Figure 1. It is clear that in all domains (except MAPPO on Spread) with either centralized training or decentralized training, smaller delays, which correspond to sharing more information will lead to faster convergence, higher final performance, and reduced training variance. For decentralized training where coordination is absent, sharing information could be more useful.

	Boxpushing			**Dectiger**		
Horizon	Ours	FM-E	RNN-E	Ours	FM-E	RNN-E
3	62.78	64.22	8.40	13.06	-6.0	-6.0
4	81.44	77.80	9.10	20.89	-4.76	-7.00
5	98.73	96.40	21.78	27.95	-6.37	-10.04
6	98.76	94.61	94.36	36.03	-7.99	-11.90
7	145.35	138.44	132.70	37.72	-7.99	-13.92

Table 1. Final evaluation rewards of our methods compared with methods FM-E and RNN-E in (Mao et al., 2020).

Validating implementability and performance. To further validate the tractability of our approaches, we test our learning algorithm on two popular partially observable benchmarks Dectiger (Nair et al., 2003) and Boxpushing (Seuken & Zilberstein, 2012). Furthermore, for scalability and compatibility with popular deep RL algorithms, we fit the transition using neural networks instead of the counting methods adopted in Algorithm 7. Both our algorithm and baselines are trained with 80000 time steps. We compare our approaches with two baselines, FM-E and RNN-E, which are also common information-based approaches in (Mao et al., 2020). The final rewards are reported in Table 1. In both domains with various horizons, our methods consistently outperform the baselines.

Acknowledgement

The authors would like to thank the anonymous reviewers for the helpful comments. K.Z. acknowledges support from Simons-Berkeley Research Fellowship and Northrop Grumman – Maryland Seed Grant Program.

References

Altman, E., Kambley, V., and Silva, A. Stochastic games with one step delay sharing information pattern with application to power control. In *2009 International Conference on Game Theory for Networks*, pp. 124–129. IEEE, 2009.

Awerbuch, B. and Kleinberg, R. Online linear optimization and adaptive routing. *Journal of Computer and System Sciences*, 74(1):97–114, 2008.

Azizzadenesheli, K., Lazaric, A., and Anandkumar, A. Reinforcement learning of POMDPs using spectral methods. In *Conference on Learning Theory*, pp. 193–256. PMLR, 2016.

Bai, Y., Jin, C., and Yu, T. Near-optimal reinforcement learning with self-play. *Advances in Neural Information Processing Systems*, 33, 2020.

Berner, C., Brockman, G., Chan, B., Cheung, V., Dębiak, P., Dennison, C., Farhi, D., Fischer, Q., Hashme, S., Hesse, C., et al. Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680*, 2019.

Bernstein, D. S., Givan, R., Immerman, N., and Zilberstein, S. The complexity of decentralized control of markov decision processes. *Mathematics of operations research*, 27(4):819–840, 2002.

Busoniu, L., Babuska, R., De Schutter, B., et al. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, 38(2):156–172, 2008.

Canonne, C. L. A short note on learning discrete distributions. *arXiv preprint arXiv:2002.11457*, 2020.

Chen, X., Deng, X., and Teng, S.-H. Settling the complexity of computing two-player Nash equilibria. *Journal of the ACM*, 56(3):14, 2009.

Daskalakis, C., Goldberg, P. W., and Papadimitriou, C. H. The complexity of computing a Nash equilibrium. *SIAM Journal on Computing*, 39(1):195–259, 2009.

Daskalakis, C., Deckelbaum, A., and Kim, A. Near-optimal no-regret algorithms for zero-sum games. In *Proceedings of the twenty-second annual ACM-SIAM symposium on Discrete Algorithms*, pp. 235–254. SIAM, 2011.

Daskalakis, C., Foster, D. J., and Golowich, N. Independent policy gradient methods for competitive reinforcement learning. In *Advances in Neural Information Processing Systems*, 2020.

Daskalakis, C., Golowich, N., and Zhang, K. The complexity of Markov equilibrium in stochastic games. *arXiv preprint arXiv:2204.03991*, 2022.

Ding, D., Wei, C.-Y., Zhang, K., and Jovanovic, M. Independent policy gradient for large-scale markov potential games: Sharper rates, function approximation, and game-agnostic convergence. In *International Conference on Machine Learning*, pp. 5166–5220. PMLR, 2022.

Efroni, Y., Jin, C., Krishnamurthy, A., and Miryoosefi, S. Provable reinforcement learning with a short-term memory. *arXiv preprint arXiv:2202.03983*, 2022.

Emery-Montemerlo, R., Gordon, G., Schneider, J., and Thrun, S. Approximate solutions for partially observable stochastic games with common payoffs. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems, 2004. AAMAS 2004.*, pp. 136–143. IEEE, 2004.

Foerster, J., Farquhar, G., Afouras, T., Nardelli, N., and Whiteson, S. Counterfactual multi-agent policy gradients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Golowich, N., Moitra, A., and Rohatgi, D. Learning in observable POMDPs, without computationally intractable oracles. In *Advances in Neural Information Processing Systems*, 2022a.

Golowich, N., Moitra, A., and Rohatgi, D. Planning in observable pomdps in quasipolynomial time. *arXiv preprint arXiv:2201.04735*, 2022b.

Gong, S., Shen, J., and Du, L. Constrained optimization and distributed computation based car following control of a connected and autonomous vehicle platoon. *Transportation Research Part B: Methodological*, 94:314–334, 2016.

Gordon, G. J., Greenwald, A., and Marks, C. No-regret learning in convex games. In *Proceedings of the 25th international conference on Machine learning*, pp. 360–367, 2008.

Gupta, A., Nayyar, A., Langbort, C., and Basar, T. Common information based markov perfect equilibria for linear-gaussian games with asymmetric information. *SIAM Journal on Control and Optimization*, 52(5):3228–3260, 2014.

Hansen, E. A., Bernstein, D. S., and Zilberstein, S. Dynamic programming for partially observable stochastic games. In *AAAI*, volume 4, pp. 709–715, 2004.

Hernandez-Leal, P., Kartal, B., and Taylor, M. E. A survey and critique of multiagent deep reinforcement learning. *Autonomous Agents and Multi-Agent Systems*, 33(6):750–797, 2019.

Ho, Y.-C. Team decision theory and information structures. *Proceedings of the IEEE*, 68(6):644–654, 1980.

Horák, K., Bošanskỳ, B., and Pěchouček, M. Heuristic search value iteration for one-sided partially observable stochastic games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

Jin, C., Kakade, S., Krishnamurthy, A., and Liu, Q. Sample-efficient reinforcement learning of undercomplete POMDPs. *Advances in Neural Information Processing Systems*, 33:18530–18539, 2020.

Jin, C., Liu, Q., Wang, Y., and Yu, T. V-learning–a simple, efficient, decentralized algorithm for multiagent rl. *arXiv preprint arXiv:2110.14555*, 2021.

Kaelbling, L. P., Littman, M. L., and Cassandra, A. R. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134, 1998.

Kao, H. and Subramanian, V. Common information based approximate state representations in multi-agent reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 6947–6967. PMLR, 2022.

Kozuno, T., Ménard, P., Munos, R., and Valko, M. Learning in two-player zero-sum partially observable Markov games with perfect recall. *Advances in Neural Information Processing Systems*, 34:11987–11998, 2021.

Krishnamurthy, A., Agarwal, A., and Langford, J. Pac reinforcement learning with rich observations. *Advances in Neural Information Processing Systems*, 29, 2016.

Kumar, A. and Zilberstein, S. Dynamic programming approximations for partially observable stochastic games. 2009.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Leonardos, S., Overman, W., Panageas, I., and Piliouras, G. Global convergence of multi-agent policy gradient in Markov potential games. In *International Conference on Learning Representations*, 2022.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. Continuous control with deep reinforcement learning. In *International Conference on Learning Representations*, 2016.

Liu, Q., Yu, T., Bai, Y., and Jin, C. A sharp analysis of model-based reinforcement learning with self-play. *arXiv preprint arXiv:2010.01604*, 2020.

Liu, Q., Yu, T., Bai, Y., and Jin, C. A sharp analysis of model-based reinforcement learning with self-play. In *International Conference on Machine Learning*, pp. 7001–7010. PMLR, 2021.

Liu, Q., Chung, A., Szepesvari, C., and Jin, C. When is partially observable reinforcement learning not scary? In *Conference on Learning Theory*, pp. 5175–5220, 2022a.

Liu, Q., Szepesvári, C., and Jin, C. Sample-efficient reinforcement learning of partially observable Markov games. In *Advances in Neural Information Processing Systems*, 2022b.

Long, P., Fan, T., Liao, X., Liu, W., Zhang, H., and Pan, J. Towards optimally decentralized multi-robot collision avoidance via deep reinforcement learning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6252–6259. IEEE, 2018.

Lowe, R., Wu, Y. I., Tamar, A., Harb, J., Pieter Abbeel, O., and Mordatch, I. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in Neural Information Processing Systems*, 30, 2017.

Lusena, C., Goldsmith, J., and Mundhenk, M. Nonapproximability results for partially observable markov decision processes. *Journal of artificial intelligence research*, 14: 83–103, 2001.

Mao, W., Zhang, K., Miehling, E., and Başar, T. Information state embedding in partially observable cooperative multi-agent reinforcement learning. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pp. 6124–6131. IEEE, 2020.

Mao, W., Yang, L., Zhang, K., and Basar, T. On improving model-free algorithms for decentralized multi-agent reinforcement learning. In *International Conference on Machine Learning*, pp. 15007–15049. PMLR, 2022.

Mundhenk, M., Goldsmith, J., Lusena, C., and Allender, E. Complexity of finite-horizon Markov decision process problems. *Journal of the ACM (JACM)*, 47(4):681–720, 2000.

Nair, R., Tambe, M., Yokoo, M., Pynadath, D., and Marsella, S. Taming decentralized pomdps: Towards efficient policy computation for multiagent settings. In *IJCAI*, volume 3, pp. 705–711, 2003.

Nayyar, A., Mahajan, A., and Teneketzis, D. Optimal control strategies in delayed sharing information structures. *IEEE Transactions on Automatic Control*, 56(7):1606–1620, 2010.

Nayyar, A., Gupta, A., Langbort, C., and Başar, T. Common information based markov perfect equilibria for stochastic games with asymmetric information: Finite games. *IEEE Transactions on Automatic Control*, 59(3):555–570, 2013a.

Nayyar, A., Mahajan, A., and Teneketzis, D. Decentralized stochastic control with partial history sharing: A common information approach. *IEEE Transactions on Automatic Control*, 58(7):1644–1658, 2013b.

Oliehoek, F. A. and Amato, C. *A Concise Introduction to Decentralized POMDPs*, volume 1. Springer, 2016.

Ouyang, Y., Tavafoghi, H., and Teneketzis, D. Dynamic games with asymmetric information: Common information based perfect Bayesian equilibria and sequential decomposition. *IEEE Transactions on Automatic Control*, 62(1):222–237, 2016.

Papadimitriou, C. H. and Tsitsiklis, J. N. The complexity of markov decision processes. *Mathematics of operations research*, 12(3):441–450, 1987.

Pathak, A., Pucha, H., Zhang, Y., Hu, Y. C., and Mao, Z. M. A measurement study of internet delay asymmetry. In *Passive and Active Network Measurement: 9th International Conference, PAM 2008, Cleveland, OH, USA, April 29-30, 2008. Proceedings 9*, pp. 182–191. Springer, 2008.

Rabinovich, Z., Goldman, C. V., and Rosenschein, J. S. The complexity of multiagent systems: The price of silence. In *Proceedings of the second international joint conference on Autonomous agents and multiagent systems*, pp. 1102–1103, 2003.

Radner, R. Team decision problems. *The Annals of Mathematical Statistics*, 33(3):857–881, 1962.

Rashid, T., Samvelyan, M., De Witt, C. S., Farquhar, G., Foerster, J., and Whiteson, S. Monotonic value function factorisation for deep multi-agent reinforcement learning. *The Journal of Machine Learning Research*, 21(1):7234–7284, 2020.

Roughgarden, T. Algorithmic game theory. *Communications of the ACM*, 53(7):78–86, 2010.

Sallab, A. E., Abdou, M., Perot, E., and Yogamani, S. Deep reinforcement learning framework for autonomous driving. *Electronic Imaging*, 2017(19):70–76, 2017.

Seuken, S. and Zilberstein, S. Improved memory-bounded dynamic programming for decentralized pomdps. *arXiv preprint arXiv:1206.5295*, 2012.

Shalev-Shwartz, S., Shammah, S., and Shashua, A. Safe, multi-agent, reinforcement learning for autonomous driving. *arXiv preprint arXiv:1610.03295*, 2016.

Shapley, L. S. Stochastic games. *Proceedings of the National Academy of Sciences*, 39(10):1095–1100, 1953.

Shi, J., Wang, G., and Xiong, J. Leader–follower stochastic differential game with asymmetric information and applications. *Automatica*, 63:60–73, 2016.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. Mastering the game of Go without human knowledge. *Nature*, 550(7676):354–359, 2017.

Song, Z., Mei, S., and Bai, Y. When can we learn general-sum Markov games with a large number of players sample-efficiently? *arXiv preprint arXiv:2110.04184*, 2021.

Subramanian, J., Sinha, A., Seraj, R., and Mahajan, A. Approximate information state for approximate planning and reinforcement learning in partially observed systems. *J. Mach. Learn. Res.*, 23:12–1, 2022.

Sunehag, P., Lever, G., Gruslys, A., Czarnecki, W. M., Zambaldi, V., Jaderberg, M., Lanctot, M., Sonnerat, N., Leibo, J. Z., Tuyls, K., et al. Value-decomposition networks for cooperative multi-agent learning based on team reward. In *International Conference on Autonomous Agents and Multi-Agent Systems*, pp. 2085–2087, 2018.

Tavafoghi, H., Ouyang, Y., and Teneketzis, D. On stochastic dynamic games with delayed sharing information structure. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pp. 7002–7009. IEEE, 2016.

Tsitsiklis, J. and Athans, M. On the complexity of decentralized decision making and detection problems. *IEEE Transactions on Automatic Control*, 30(5):440–446, 1985.

Uehara, M., Sekhari, A., Lee, J. D., Kallus, N., and Sun, W. Provably efficient reinforcement learning in partially observable dynamical systems. *arXiv preprint arXiv:2206.12020*, 2022.

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., Choi, D. H., Powell, R., Ewalds, T., Georgiev, P., et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575 (7782):350–354, 2019.

Wang, L., Cai, Q., Yang, Z., and Wang, Z. Embed to control partially observed systems: Representation learning with provable sample efficiency. *arXiv preprint arXiv:2205.13476*, 2022.

Witsenhausen, H. S. A counterexample in stochastic optimum control. *SIAM Journal on Control*, 6(1):131–147, 1968.

Witsenhausen, H. S. Separation of estimation and control for discrete time systems. *Proceedings of the IEEE*, 59 (11):1557–1566, 1971.

Xie, Q., Chen, Y., Wang, Z., and Yang, Z. Learning zero-sum simultaneous-move markov games using function approximation and correlated equilibrium. In *Conference on learning theory*, pp. 3674–3682. PMLR, 2020.

Yu, C., Velu, A., Vinitsky, E., Wang, Y., Bayen, A., and Wu, Y. The surprising effectiveness of ppo in cooperative, multi-agent games. *arXiv preprint arXiv:2103.01955*, 2021.

Zhan, W., Uehara, M., Sun, W., and Lee, J. D. PAC reinforcement learning for predictive state representations. *arXiv preprint arXiv:2207.05738*, 2022.

Zhang, K., Kakade, S. M., Başar, T., and Yang, L. F. Model-based multi-agent RL in zero-sum Markov games with near-optimal sample complexity. *arXiv preprint arXiv:2007.07461*, 2020.

Zhang, K., Yang, Z., and Başar, T. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of Reinforcement Learning and Control*, pp. 321–384, 2021a.

Zhang, K., Zhang, X., Hu, B., and Basar, T. Derivative-free policy optimization for linear risk-sensitive and robust control design: Implicit regularization and sample complexity. *Advances in Neural Information Processing Systems*, 34:2949–2964, 2021b.

Zhang, R., Ren, Z., and Li, N. Gradient play in stochastic games: stationary points, convergence, and sample complexity. *arXiv preprint arXiv:2106.00198*, 2021c.

Zinkevich, M., Johanson, M., Bowling, M., and Piccione, C. Regret minimization in games with incomplete information. In *Advances in Neural Information Processing Systems*, pp. 1729–1736, 2007.

Supplementary Materials for "Partially Observable Multi-agent RL with (Quasi-)Efficiency: The Blessing of Information Sharing"

A. Related Work

Decentralized stochastic control and decision-making. Decentralized stochastic control and decision-making are known to have unique challenges, compared to the single-agent counterpart, since the seminal works (Witsenhausen, 1968; Tsitsiklis & Athans, 1985). In particular, (Tsitsiklis & Athans, 1985) shows that variations of the classical "team decision problem" can be `NP-hard`. Later, (Bernstein et al., 2002) shows that planning in Dec-POMDPs, a special class of POSGs with identical reward functions among agents, can be `NEXP-hard` in finding the team-optimal solution. (Hansen et al., 2004) provides a popular POSG planning algorithm, though without any complexity guarantees. There also exist other approximate/heuristic algorithms for solving POSGs (Emery-Montemerlo et al., 2004; Kumar & Zilberstein, 2009; Horák et al., 2017).

Information sharing in theory and practice. The idea of information-sharing has been explored in decentralized stochastic control (Witsenhausen, 1971; Nayyar et al., 2010; 2013b), as well as stochastic games with asymmetric information (Nayyar et al., 2013a; Gupta et al., 2014; Ouyang et al., 2016). The common-information-based approach in the seminal works (Nayyar et al., 2013a;b) provides significant inspiration for our work. However, no computation nor sample complexities of algorithms were discussed in these works. On the other hand, information-sharing has become a normal practice in empirical MARL, especially recently in deep MARL (Lowe et al., 2017; Foerster et al., 2018; Sunehag et al., 2018; Rashid et al., 2020). The sharing was instantiated via so-called *centralized training*, where all agents' information was shared in training. Centralized training with shared information has been shown to significantly improve the learning efficiency. One caveat is that these empirical works also popularize the *centralized-training-decentralized-execution* paradigm, while our MARL algorithms under the common-information sharing framework require sharing in both training/learning and execution.

Provable multi-agent reinforcement learning. There has been fast-growing literature on provable MARL algorithms with sample efficiency guarantees, e.g., (Bai et al., 2020; Liu et al., 2020; Zhang et al., 2020; Xie et al., 2020; Zhang et al., 2021b; Daskalakis et al., 2020; Jin et al., 2021; Song et al., 2021; Daskalakis et al., 2022; Mao et al., 2022). However, these works have exclusively focused on the fully observable Markov/stochastic games. The only MARL algorithms under partial observability that enjoy finite-sample guarantees, to the best of our knowledge, are those in (Liu et al., 2022b; Kozuno et al., 2021). However, the algorithms in (Kozuno et al., 2021) only apply to POSGs with certain tree-structured transitions, while those in (Liu et al., 2022b) require computationally intractable oracles.

RL in partially observable environments. It is known that in general, even planning in single-agent POMDPs can be `PSPACE-complete` (Papadimitriou & Tsitsiklis, 1987) and thus computationally hard. Statistically, learning POMDPs can also be hard in general (Krishnamurthy et al., 2016; Jin et al., 2020). There has thus been a growing literature on RL in POMDPs with additional assumptions, e.g., (Azizzadenesheli et al., 2016; Jin et al., 2020; Liu et al., 2022a; Wang et al., 2022; Zhan et al., 2022). However, these works only focus on statistical efficiency, and the algorithms usually require computationally intractable oracles. More recently, (Golowich et al., 2022b) has identified the condition of γ-observability in POMDPs, and has shown that quasi-polynomial-time-complexity planning algorithm exists for solving such POMDPs. Subsequently, (Golowich et al., 2022a) has developed an RL algorithm based on the planning one in (Golowich et al., 2022b), which is both sample and computation quasi-efficient.

B. Additional Definitions and Examples

B.1. Belief states

In such partially observable games, each agent cannot know the underlying state but could infer the underlying distribution of states through the observations and actions. Following the convention in POMDPs, we call such distributions as the belief

states. Such posterior distributions over states can be updated whenever the agent receives new observations and actions. Formally, we define the belief update as:

Definition 8 (Belief state update)**.** For each $h \in [H+1]$, the Bayes operator (with respect to the joint observation) $B_h : \Delta(\mathcal{S}) \times \mathcal{O} \to \Delta(\mathcal{S})$ is defined for $b \in \Delta(\mathcal{S})$, and $y \in \mathcal{O}$ by:

$$B_h(b; y)(x) = \frac{\mathbb{O}_h(y \mid x) b(x)}{\sum_{z \in \mathcal{S}} \mathbb{O}_h(y \mid z) b(z)}.$$

Similarly, for each $h \in [H], i \in [n]$, we define the Bayes operator with respect to individual observations $B_{i,h} : \Delta(\mathcal{S}) \times \mathcal{O}_i \to \Delta(\mathcal{S})$ by:

$$B_{i,h}(b; y)(x) = \frac{\mathbb{O}_{i,h}(y \mid x) b(x)}{\sum_{z \in \mathcal{S}} \mathbb{O}_{i,h}(y \mid z) b(z)}.$$

For each $h \in [H]$, the belief update operator $U_h : \Delta(\mathcal{S}) \times \mathcal{A} \times \mathcal{O} \to \Delta(\mathcal{S})$, is defined by

$$U_h(b; a, y) = B_{h+1}\left(\mathbb{T}_h(a) \cdot b; y\right),$$

where $\mathbb{T}_h(a) \cdot b$ represents the matrix multiplication. We use the notation $\boldsymbol{b}_h$ to denote the belief update function, which receives a sequence of actions and observations and outputs a distribution over states at the step h: the belief state at step $h = 1$ is defined as $\boldsymbol{b}_1(\emptyset) = \mu_1$. For any $1 \le h \le H$ and any action-observation sequence $(a_{1:h-1}, o_{2:h})$, we inductively define the belief state:

$$\begin{aligned}
\boldsymbol{b}_{h+1}(a_{1:h}, o_{2:h}) &= \mathbb{T}_h(a_h) \cdot \boldsymbol{b}_h(a_{1:h-1}, o_{2:h}), \\
\boldsymbol{b}_h(a_{1:h-1}, o_{2:h}) &= B_h(\boldsymbol{b}_h(a_{1:h-1}, o_{2:h-1}); o_h).
\end{aligned}$$

Also, we slightly abuse the notation and define the belief state containing individual observations as

$$\boldsymbol{b}_h(a_{1:h-1}, o_{2:h-1}, o_{i,h}) = B_{i,h}(\boldsymbol{b}_h(a_{1:h-1}, o_{2:h-1}); o_{i,h}).$$

We also define the approximate belief update using the most recent L-step history. For $1 \le h \le H$, we follow the notation of (Golowich et al., 2022b) and define

$$\boldsymbol{b}_h^{\mathrm{apx},\mathcal{G}}(\emptyset; D) = \begin{cases} \mu_1 & \text{if } h = 1 \\ D & \text{otherwise}, \end{cases}$$

where $D \in \Delta(\mathcal{S})$ is the prior for the approximate belief update. Then for any $1 \le h - L < h \le H$ and any action-observation sequence $(a_{h-L:h-1}, o_{h-L+1:h})$, we inductively define

$$\begin{aligned}
\boldsymbol{b}_{h+1}^{\mathrm{apx},\mathcal{G}}(a_{h-L:h}, o_{h-L+1:h}; D) &= \mathbb{T}_h(a_h) \cdot \boldsymbol{b}_h^{\mathrm{apx},\mathcal{G}}(a_{h-L:h-1}, o_{h-L+1:h}; D), \\
\boldsymbol{b}_h^{\mathrm{apx},\mathcal{G}}(a_{h-L:h-1}, o_{h-L+1:h}; D) &= B_h(\boldsymbol{b}_h^{\mathrm{apx},\mathcal{G}}(a_{h-L:h-1}, o_{h-L+1:h-1}; D); o_h).
\end{aligned}$$

For the remainder of our paper, we shall use the important initialization for the approximate belief, which are defined as

$$\boldsymbol{b}'_h(a_{h-L:h-1}, o_{h-L+1:h}) := \boldsymbol{b}_h^{\mathrm{apx},\mathcal{G}}(a_{h-L:h-1}, o_{h-L+1:h}; \mathrm{Unif}(\mathcal{S})).$$

B.2. Additional definitions of solution concepts

Definition 9 (ϵ-approximate Coarse Correlated Equilibrium)**.** For any $\epsilon \ge 0$, a joint policy $\pi^\star \in \Pi$ is an ϵ-approximate Coarse Correlated Equilibrium of the POSG $\mathcal{G}$ if:

$$\text{CCE-gap}(\pi^\star) := \max_i \left(\max_{\pi'_i \in \Pi_i} V_{i,1}^{\pi'_i \times \pi^\star_{-i}, \mathcal{G}}(\emptyset) - V_{i,1}^{\pi^\star, \mathcal{G}}(\emptyset) \right) \le \epsilon.$$

Definition 10 (ϵ-approximate Correlated Equilibrium)**.** For any $\epsilon \geq 0$, a joint policy $\pi^\star \in \Pi$ is an ϵ-approximate Correlated Equilibrium of the POSG $\mathcal{G}$ if:

$$\text{CE-gap}(\pi^\star) := \max_i \left(\max_{\phi_i} V_{i,1}^{(\phi_i \diamond \pi_i^\star) \odot \pi_{-i}^\star, \mathcal{G}}(\emptyset) - V_{i,1}^{\pi^\star, \mathcal{G}}(\emptyset) \right) \leq \epsilon,$$

where ϕ_i is called *strategy modification* and $\phi_i = \{\phi_{i,h,c_h,p_{i,h}}\}_{h,c_h,p_{i,h}}$, with each $\phi_{i,h,c_h,p_{i,h}} : \mathcal{A}_i \to \mathcal{A}_i$ being a mapping from the action set to itself. The space of ϕ_i is denoted as Φ_i. The composition $\phi_i \diamond \pi_i$ will work as follows: at the step h, when the agent i is given c_h and $p_{i,h}$, the action chosen to be $(a_{1,h}, \cdots, a_{i,h}, \cdots, a_{n,h})$ will be modified to $(a_{1,h}, \cdots, \phi_{i,h,c_h,p_{i,h}}(a_{i,h}), \cdots, a_{n,h})$. Note this definition follows the definition in (Song et al., 2021; Liu et al., 2021; Jin et al., 2021) when there exists common information and is a natural generalization from the normal-form game case (Roughgarden, 2010).

B.3. Additional definitions of value functions and policies

First, we define the prescription-value function in the POSG $\mathcal{G}$ below as a generalization of action-value function in Markov game.

Definition 11 (Prescription-value function)**.** At step h, given the common information c_h, joint policies $\pi = \{\pi_i\}_{i=1}^n$, and prescriptions $\{\gamma_{i,h}\}_{i=1}^n$, the prescription-value function of the i^{th} agent is defined as:

$$Q_{i,h}^{\pi,\mathcal{G}}(c_h, \{\gamma_{j,h}\}_{j\in[n]}) := \mathbb{E}_\pi^{\mathcal{G}}[r_{i,h+1}(o_{h+1}) + V_{i,h+1}^{\pi,\mathcal{G}}(c_{h+1}) | c_h, \{\gamma_{j,h}\}_{j\in[n]}],$$

where the prescription $\gamma_{i,h} \in \Delta(\mathcal{A}_i)^{P_{i,h}}$ replaces the partial function $\pi_{i,h}(\cdot | \omega_{i,h}, c_h, \cdot)$ in the value function.

This prescription-value function indicates the expected return for the i^{th} agent when all agents firstly adopt the prescriptions $\{\gamma_{j,h}\}_{j\in[n]}$ and then follow the policy π.

With the expected approximate common information model $\mathcal{M}$ defined in Definition 5, we can define the value function and policy under $\mathcal{M}$ accordingly as follows.

Definition 12 (Value function and policy under $\mathcal{M}$)**.** Given an approximate common information model $\mathcal{M}$. For any policy $\pi \in \Pi$, for each $i \in [n], h \in [H]$, we define the value function as

$$V_{i,h}^{\pi,\mathcal{M}}(c_h) = \mathbb{E}_{\{\omega_{j,h}\}_{j\in[n]}} \mathbb{E}^{\mathcal{M}}[\widehat{r}_{i,h+1}(o_{h+1}) + V_{i,h+1}^{\pi,\mathcal{M}}(c_{h+1}) | \widehat{c}_h, \{\pi_{j,h}(\cdot | \omega_{j,h}, c_h, \cdot)_{j\in[n]}\}]. \tag{B.1}$$

For any $c_{H+1} \in \mathcal{C}_{H+1}$, we define $V_{i,H+1}^{\pi,\mathcal{M}}(c_{H+1}) = 0$. Furthermore, for a policy $\widehat{\pi}$ whose $\widehat{\pi}_{i,h} : \Omega_h \times \mathcal{P}_{i,h} \times \widehat{\mathcal{C}}_h \to \Delta(\mathcal{A}_i)$ takes *approximate* instead of the exact common information as the input, we define

$$V_{i,h}^{\widehat{\pi},\mathcal{M}}(\widehat{c}_h) = \mathbb{E}_{\{\omega_{j,h}\}_{j\in[n]}} \mathbb{E}^{\mathcal{M}}[\widehat{r}_{i,h+1}(o_{h+1}) + V_{i,h+1}^{\widehat{\pi},\mathcal{M}}(\widehat{c}_{h+1}) | \widehat{c}_h, \{\widehat{\pi}_{j,h}(\cdot | \omega_{j,h}, \widehat{c}_h, \cdot)_{j\in[n]}\}], \tag{B.2}$$

where similarly, for each $\widehat{c}_{H+1} \in \widehat{\mathcal{C}}_{H+1}$, we define $V_{i,H+1}^{\widehat{\pi},\mathcal{M}}(\widehat{c}_{H+1}) = 0$. With a slight abuse of notation, sometimes $\widehat{\pi}_{i,h}$ may also take $c_h \in \mathcal{C}_h$ as input and thus $\widehat{\pi} \in \Pi$. In this case, when $\mathcal{M}$ and the corresponding compression function Compress_h are clear from the context, it means $\widehat{\pi}_{i,h}(\cdot | \cdot, c_h, \cdot) := \widehat{\pi}_{i,h}(\cdot | \cdot, \text{Compress}_h(c_h), \cdot)$. Accordingly, in this case, the definitions of $V_{i,h}^{\widehat{\pi},\mathcal{G}}(c_h)$ and $V_{i,h}^{\widehat{\pi},\mathcal{M}}(c_h)$ follows from Definition 1 and Equation (B.1), respectively.

B.4. More examples of information sharing

Example 4 (Information sharing with one-directional-one-step delay)**.** Similar to the previous cases, we also assume there are 2 players for convenience. Similar to the one-step delay case, we consider the situation where all observations of the player 1 are available to player 2, while the observations of player 2 are available to player 1 with one-step delay. All past actions are available to both players. That is, in this case, $c_h = \{o_{1,2:h}, o_{2,2:h-1}, a_{1:h-1}\}$, and player 1 has no private information, i.e., $p_{2,h} = \emptyset$, and player 2 has private information $p_{2,h} = \{o_{2,h}\}$.

Example 5 (Uncontrolled state process)**.** Consider the case where the state transition does not depend on the actions, that is, $\mathbb{T}_h(\cdot \mid s_h, a_h) = \mathbb{T}_h(\cdot \mid s_h, a'_h)$ for any s_h, a_h, a'_h, h. Note here the evolution of common and private information does not need include the actions anymore since doing so does not lose any optimality. Meanwhile, different agents are still coupled through the joint cost. An example of this case is the information structure where controllers share their observations with a delay of $d \geq 1$ time steps. In this case, the common information is $c_h = \{o_{2:h-d}\}$ and the private information is $p_{i,h} = \{o_{i,h-d+1:h}\}$.

B.5. Model-belief consistency

Definition 13. We say the approximate common information model $\mathcal{M}$ is *consistent with* some belief $\{\mathbb{P}_h^{\mathcal{M},c}(s_h, p_h|\widehat{c}_h)\}_{h\in[H]}$ if it satisfies the following for any $i \in [n]$, $h \in [H]$,

$$\mathbb{P}_h^{\mathcal{M},z}(z_{h+1}|\widehat{c}_h, \gamma_h) = \sum_{s_h} \sum_{p_h,a_h,o_{h+1}:\chi_{h+1}(p_h,a_h,o_{h+1})=z_{h+1}} \mathbb{P}_h^{\mathcal{M},c}(s_h, p_h|\widehat{c}_h) \quad \Pi_{j=1}^n \gamma_{j,h}(a_{j,h}|p_{j,h}) \sum_{s_{h+1}} \mathbb{T}_h(s_{h+1}|s_h, a_h)\mathbb{O}_{h+1}(o_{h+1}|s_{h+1}), \tag{B.3}$$

$$\mathbb{P}_h^{\mathcal{M},o}(o_{h+1}|\widehat{c}_h, \gamma_h) = \sum_{s_h,p_h,a_h} \mathbb{P}_h^{\mathcal{M},c}(s_h, p_h|\widehat{c}_h) \quad \Pi_{j=1}^n \gamma_{j,h}(a_{j,h}|p_{j,h}) \sum_{s_{h+1}} \mathbb{T}_h(s_{h+1}|s_h, a_h)\mathbb{O}_{h+1}(o_{h+1}|s_{h+1}). \tag{B.4}$$

C. Collection of Algorithm Pseudocodes

Here we collect both our planning and learning algorithms as in Algorithms 1, 2, 3, 4, 5, 6. For a high-level overview of our algorithmic framework, we refer to Figure 2.

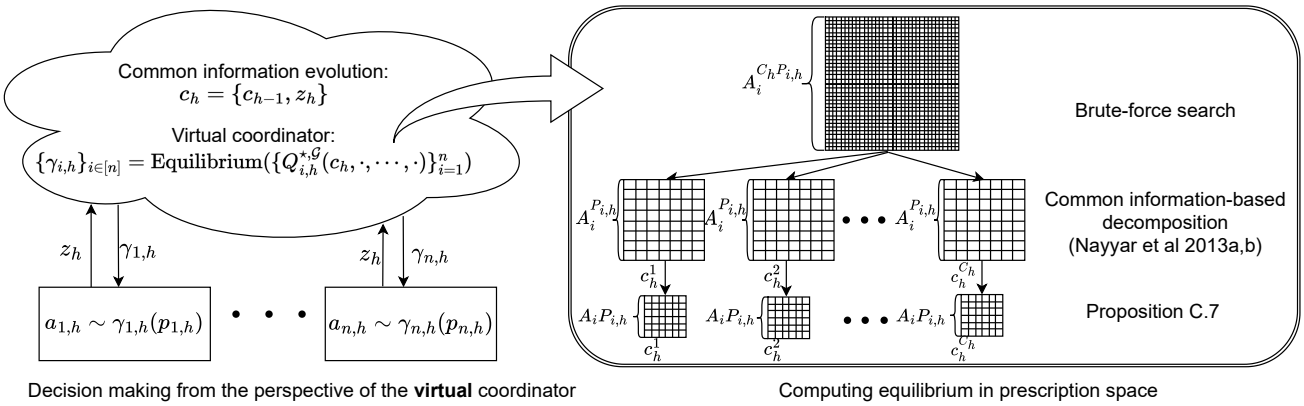

Figure 2. An overview of our algorithmic framework. The left part of the figure shows that there is a virtual coordinator collecting the information shared among agents. Based on the common information c_h, it will compute an equilibrium in the prescription space and assign it to all the agents. The right part shows the computation of equilibrium. Let's take the example of $A_i = 2$, $P_{i,h} = 3$, $C_h = 2$. If we search over all deterministic prescriptions, the corresponding matrix game will have the size of $A_i^{C_h P_{i,h}} = 64$. Then (Nayyar et al., 2013a;b) propose the common information-based decomposition and solves C_h games of smaller size. However, (Nayyar et al., 2013b) treats each deterministic prescription as an action and the size of each sub-game will be $A_i^{P_{i,h}} = 8$. Furthermore, Proposition 9 shows that we can reformulate each sub-game as a game whose payoff is multi-linear with respect to each agent's prescription, and whose dimensionality is $A_i P_{i,h} = 6$.

Algorithm 1 Value iteration with common information

Input: $\mathcal{G}, \epsilon_e$
for each $i \in [n]$ and c_{H+1} **do**
  $V^{\star,\mathcal{G}}_{i,H+1}(c_{H+1}) \leftarrow 0$
end for
for $h = H, \cdots, 1$ **do**
  for each c_h **do**
    Define $Q^{\star,\mathcal{G}}_{i,h}(c_h, \gamma_{1,h}, \cdots, \gamma_{n,h}) := \mathbb{E}_{s_h,p_h\sim\mathbb{P}^{\mathcal{G}}_h(\cdot,\cdot|c_h)}\mathbb{E}_{\{a_{j,h}\sim\gamma_{j,h}(\cdot|p_{j,h})\}_{j\in[n]}}\mathbb{E}_{o_{h+1}\sim\mathbb{O}^\top_{h+1}\mathbb{T}_h(\cdot|s_h,a_h)}[r_{i,h+1}(o_{h+1}) + V^{\star,\mathcal{G}}_{i,h+1}(c_{h+1})]$
    $\left(\pi^\star_{1,h}(\cdot|\cdot,c_h,\cdot),\cdots,\pi^\star_{n,h}(\cdot|\cdot,c_h,\cdot)\right) \leftarrow \text{NE/CE/CCE}(\{Q^{\star,\mathcal{G}}_{i,h}(c_h,\cdot,\cdots,\cdot)\}^n_{i=1}, \epsilon_e)$ // we refer the implementation to §E.2
    for each $i \in [n]$ **do**
      $V^{\star,\mathcal{G}}_{i,h}(c_h) \leftarrow \mathbb{E}_{\{\omega_{j,h}\}_{j\in[n]}}\mathbb{E}_{s_h,p_h\sim\mathbb{P}^{\mathcal{G}}_h(\cdot,\cdot|c_h)}\mathbb{E}_{\{a_{j,h}\sim\pi^\star_{j,h}(\cdot|\omega_{j,h},c_h,p_{j,h})\}_{j\in[n]}}\mathbb{E}_{o_{h+1}\sim\mathbb{O}^\top_{h+1}\mathbb{T}_h(\cdot|s_h,a_h)}[r_{i,h+1}(o_{h+1}) + V^{\star,\mathcal{G}}_{i,h+1}(c_{h+1})]$
    end for
  end for
end for
return $\pi^\star$

Algorithm 2 $\texttt{BR}(\mathcal{G}, \pi, i, \epsilon_e)$: **B**est **R**esponse for agent i

Input: $\mathcal{G}, \pi, i, \epsilon_e$
$V^{\star,\mathcal{G}}_{i,H+1}(c_{H+1}) \leftarrow 0$ for all c_{H+1}
for $h = H, \cdots, 1$ **do**
  for each c_h **do**
    Define $Q^{\star,\mathcal{G}}_{i,h}(c_h, \gamma_{1,h}, \cdots, \gamma_{n,h}) := \mathbb{E}_{s_h, p_h \sim \mathbb{P}^{\mathcal{G}}_h(\cdot,\cdot|c_h)} \mathbb{E}_{\{a_{j,h} \sim \gamma_{j,h}(\cdot|p_{j,h})\}_{j\in[n]}} \mathbb{E}_{o_{h+1}\sim \mathbb{O}^{\top}_{h+1}\mathbb{T}_h(\cdot|s_h,a_h)}[r_{i,h+1}(o_{h+1}) + V^{\star,\mathcal{G}}_{i,h+1}(c_{h+1})]$
    $\pi^{\star}_{i,h}(\cdot|\cdot, c_h, \cdot) \leftarrow \text{NE/CE/CCE-BR}(Q^{\star,\mathcal{G}}_{i,h}(c_h, \cdot, \cdots, \cdot), \{\pi_{j,h}(\cdot|\cdot, c_h, \cdot)\}_{j\in[n]}, i, \epsilon_e)$ // we refer the implementation to §E.2
    $V^{\star,\mathcal{G}}_{i,h}(c_h) \leftarrow \mathbb{E}_{\{\omega_{j,h}\}_{j\in[n]}} \mathbb{E}_{s_h, p_h \sim \mathbb{P}^{\mathcal{G}}_h(\cdot,\cdot|c_h)} \mathbb{E}_{a_{i,h}\sim\pi^{\star}_{i,h}(\cdot|\omega_{i,h}, c_h, p_{i,h}), a_{-i,h}\sim\pi_{-i,h}(\cdot|\omega_{-i,h}, c_h, p_{-i,h})} \mathbb{E}_{o_{h+1}\sim\mathbb{O}^{\top}_{h+1}\mathbb{T}_h(\cdot|s_h,a_h)}[r_{i,h+1}(o_{h+1}) + V^{\star,\mathcal{G}}_{i,h+1}(c_{h+1})]$
  end for
end for
return $\pi^{\star}_i$

Algorithm 3 $\texttt{VIACM}(\mathcal{M}, \epsilon_e)$: **V**alue **I**teration with (expected) **A**pproximate **C**ommon-Information **M**odel

Input: $\mathcal{M}, \epsilon_e$
for each $i \in [n]$ and $\widehat{c}_{H+1}$ **do**
  $V^{\star,\mathcal{M}}_{i,H+1}(\widehat{c}_{H+1}) \leftarrow 0$
end for
for $h = H, \cdots, 1$ **do**
  for each $\widehat{c}_h$ **do**
    Define $Q^{\star,\mathcal{M}}_{i,h}(\widehat{c}_h, \gamma_{1,h}, \cdots, \gamma_{n,h}) := \mathbb{E}^{\mathcal{M}}[\widehat{r}_{i,h+1}(o_{h+1}) + V^{\star,\mathcal{M}}_{i,h+1}(\widehat{c}_{h+1}) \,|\, \widehat{c}_h, \{\gamma_{j,h}\}_{j\in[n]}]$ for any $i \in [n]$
    $\left(\widehat{\pi}^{\star}_{1,h}(\cdot|\cdot, \widehat{c}_h, \cdot), \cdots, \widehat{\pi}^{\star}_{n,h}(\cdot|\cdot, \widehat{c}_h, \cdot)\right) \leftarrow \text{NE/CE/CCE}(\{Q^{\star,\mathcal{M}}_{i,h}(\widehat{c}_h, \cdot, \cdots, \cdot)\}^n_{i=1}, \epsilon_e)$ // we refer the implementation to §E.2
    for each $i \in [n]$ **do**
      $V^{\star,\mathcal{M}}_{i,h}(\widehat{c}_h) \leftarrow \mathbb{E}_{\{\omega_{j,h}\}_{j\in[n]}} \mathbb{E}^{\mathcal{M}}[\widehat{r}_{i,h+1}(o_{h+1}) + V^{\star,\mathcal{M}}_{i,h+1}(\widehat{c}_{h+1}) \,|\, \widehat{c}_h, \{\widehat{\pi}^{\star}_{j,h}(\cdot|\omega_{j,h}, \widehat{c}_h, \cdot)\}_{j\in[n]}]$
    end for
  end for
end for
return $\widehat{\pi}^{\star}$

Algorithm 4 $\texttt{ABR}(\mathcal{M}, \widehat{\pi}, i, \epsilon_e)$: **A**pproximate **B**est **R**esponse for agent i under approximate common information model

Input: $\mathcal{M}, \widehat{\pi}, i, \epsilon_e$
$V^{\star,\mathcal{M}}_{i,H+1}(\widehat{c}_{H+1}) \leftarrow 0$ for all $\widehat{c}_{H+1}$
for $h = H, \cdots, 1$ **do**
  for each $\widehat{c}_h$ **do**
    Define $Q^{\star,\mathcal{M}}_{i,h}(\widehat{c}_h, \gamma_{1,h}, \cdots, \gamma_{n,h}) := \mathbb{E}^{\mathcal{M}}[\widehat{r}_{i,h+1}(o_{h+1}) + V^{\star,\mathcal{M}}_{i,h+1}(\widehat{c}_{h+1}) \,|\, \widehat{c}_h, \{\gamma_{j,h}\}_{j\in[n]}]$
    $\widehat{\pi}^{\star}_{i,h}(\cdot|\cdot, \widehat{c}_h, \cdot) \leftarrow \text{NE/CE/CCE-BR}(Q^{\star,\mathcal{M}}_{i,h}(\widehat{c}_h, \cdot, \cdots, \cdot), \{\widehat{\pi}_{j,h}(\cdot|\cdot, \widehat{c}_h, \cdot)\}_{j\in[n]}, i, \epsilon_e)$ // we refer the implementation to §E.2
    $V^{\star,\mathcal{M}}_{i,h}(\widehat{c}_h) \leftarrow \mathbb{E}_{\{\omega_{j,h}\}_{j\in[n]}} \mathbb{E}^{\mathcal{M}}[\widehat{r}_{i,h+1}(o_{h+1}) + V^{\star,\mathcal{M}}_{i,h+1}(\widehat{c}_{h+1}) \,|\, \widehat{c}_h, \{\widehat{\pi}^{\star}_{i,h}(\cdot|\omega_{i,h}, \widehat{c}_h, \cdot), \widehat{\pi}_{-i,h}(\cdot|\omega_{-i,h}, \widehat{c}_h, \cdot)\}]$
  end for
end for
return $\widehat{\pi}^{\star}_i$

Algorithm 5 $\texttt{Construct}(\pi^{1:H,j}, \{\widehat{r}_i^j\}_{i=1}^n, \{\widehat{\mathcal{C}}_h\}_{h\in[H+1]}, \{\widehat{\phi}_{h+1}\}_{h\in[H]}, \Gamma, \zeta_1, \zeta_2, \theta_1, \theta_2, \delta_1)$: Constructing empirical estimator $\widehat{\mathcal{M}}(\pi^{1:H})$ of $\widetilde{\mathcal{M}}(\pi^{1:H})$

Input: $\pi^{1:H,j}, \{\widehat{r}_i^j\}_{i=1}^n, \{\widehat{\mathcal{C}}_h\}_{h\in[H+1]}, \{\widehat{\phi}_{h+1}\}_{h\in[H]}, \Gamma, \zeta_1, \zeta_2, \theta_1, \theta_2, \delta_1$
for $1 \le h \le H$ **do**
  Define N_0 as in Equation (D.1).
  Draw N_0 independent trajectories by executing the policy $\pi^{h,j}$, and denote the k^{th} trajectory by $(a^k_{1:H-1}, o^k_{2:H})$, for $k \in [N_0]$.
  for each $\widehat{c}_h \in \widehat{\mathcal{C}}_h$ **do**
    Define $\varphi(p_h) := |\{k : \text{Compress}_h(f_h(a^k_{1:h-1}, o^k_{2:h})) = \widehat{c}_h, \text{ and } g_h(a^k_{1:h-1}, o^k_{2:h}) = p_h\}|$.
    Set $\mathbb{P}_h^{\widehat{\mathcal{M}}(\pi^{1:H})}(p_h|\widehat{c}_h) := \frac{\varphi(p_h)}{\sum_{p'_h}\varphi(p'_h)}$ for all $p_h \in \mathcal{P}_h$.
  end for
  for each $\widehat{c}_h \in \widehat{\mathcal{C}}_h$, $p_h \in \mathcal{P}_h$, $a_h \in \mathcal{A}$ **do**
    Define $\psi(o_{h+1}) := |\{k : \text{Compress}_h(f_h(a^k_{1:h-1}, o^k_{2:h})) = \widehat{c}_h, g_h(a^k_{1:h-1}, o^k_{2:h}) = p_h, a^k_h = a_h, \text{ and } o^k_{h+1} = o_{h+1}\}|$.
    Set $\mathbb{P}_h^{\widehat{\mathcal{M}}(\pi^{1:H})}(o_{h+1}|\widehat{c}_h, p_h, a_h) := \frac{\psi(o_{h+1})}{\sum_{o'_{h+1}}\psi(o'_{h+1})}$ for all $o_{h+1} \in \mathcal{O}$.
  end for
end for
Define for any $h \in [H]$, $\widehat{c}_h \in \widehat{\mathcal{C}}_h$, $\gamma_h \in \Gamma_h$, $o_{h+1} \in \mathcal{O}_{h+1}$, $z_{h+1} \in \mathcal{Z}_{h+1}$:

$$\mathbb{P}_h^{\widehat{\mathcal{M}}(\pi^{1:H}),z}(z_{h+1}|\widehat{c}_h, \gamma_h) \leftarrow \sum_{p_h, a_h, o_{h+1}: \chi_{h+1}(p_h, a_h, o_{h+1}) = z_{h+1}} \mathbb{P}_h^{\widehat{\mathcal{M}}(\pi^{1:H})}(p_h|\widehat{c}_h)\Pi_{i=1}^n \gamma_{i,h}(a_{i,h}|p_{i,h})\mathbb{P}_h^{\widehat{\mathcal{M}}(\pi^{1:H})}(o_{h+1}|\widehat{c}_h, p_h, a_h)$$

$$\mathbb{P}_h^{\widehat{\mathcal{M}}(\pi^{1:H}),o}(o_{h+1}|\widehat{c}_h, \gamma_h) \leftarrow \sum_{p_h, a_h} \mathbb{P}_h^{\widehat{\mathcal{M}}(\pi^{1:H})}(p_h|\widehat{c}_h)\Pi_{i=1}^n \gamma_{i,h}(a_{i,h}|p_{i,h})\mathbb{P}_h^{\widehat{\mathcal{M}}(\pi^{1:H})}(o_{h+1}|\widehat{c}_h, p_h, a_h)$$

return $\widehat{\mathcal{M}}(\pi^{1:H}) := (\{\widehat{\mathcal{C}}_h\}_{h\in[H+1]}, \{\widehat{\phi}_{h+1}\}_{h\in[H]}, \{\mathbb{P}_h^{\widehat{\mathcal{M}}(\pi^{1:H}),z}, \mathbb{P}_h^{\widehat{\mathcal{M}}(\pi^{1:H}),o}\}_{h\in[H]}, \Gamma, \{\widehat{r}_i^j\}_{i=1}^n)$

Algorithm 6 $\texttt{PoS}(\{\widehat{\mathcal{M}}(\pi^{1:H,j})\}_{j\in[m]}, \{\pi^{\star,j}\}_{j=1}^K, \epsilon_e, N_2)$: **Po**licy **S**election

Input: $\{\widehat{\mathcal{M}}(\pi^{1:H,j})\}_{j\in[m]}, \{\pi^{\star,j}\}_{j=1}^K, \epsilon_e, N_2$
for $i \in [n], j \in [K], m \in [K]$ **do**
  $\pi_i^{\star,j,m} \leftarrow \texttt{ABR}(\widehat{\mathcal{M}}(\pi^{1:H,m}), \pi^{\star,j}, i, \epsilon_e)$ //i.e., Algorithm 4
end for
for $j \in [K]$ **do**
  Execute $\pi^{\star,j}$ for N_2 trajectories and let the mean reward for player i be R_i^j
end for
for $i \in [n], j \in [K], m \in [K]$ **do**
  Execute $\pi_i^{\star,j,m} \odot \pi_{-i}^{\star,j}$ for N_2 trajectories and let the mean reward for player i be $R_i^{j,m}$
end for
$\widehat{j} \leftarrow \arg\min_j \left(\max_i \max_m (R_i^{j,m} - R_i^j)\right)$
return $\pi^{\star,\widehat{j}}$

Algorithm 7 `LACI`$(\mathcal{G}, \{\widehat{\mathcal{C}}_h\}_{h\in[H+1]}, \{\widehat{\phi}_{h+1}\}_{h\in[H]}, \Gamma, \widehat{L}, \epsilon, \delta_2, \zeta_1, \zeta_2, \theta_1, \theta_2, \delta_1, N_2, \epsilon_e)$: **L**earning with **A**pproximate **C**ommon **I**nformation

Input: $\mathcal{G}, \{\widehat{\mathcal{C}}_h\}_{h\in[H+1]}, \{\widehat{\phi}_{h+1}\}_{h\in[H]}, \Gamma, \widehat{L}, \epsilon, \delta_2, \zeta_1, \zeta_2, \theta_1, \theta_2, \delta_1, N_2, \epsilon_e$

$\{\pi^{1:H,j}\}_{j=1}^K, \{\{\widehat{r}_i^j\}_{i=1}^n\}_{j=1}^K \leftarrow$ `BaSeCAMP`$(\mathcal{G}, \widehat{L}, \epsilon, \delta_2)$ // i.e., Algorithm 3 of (Golowich et al., 2022a)

for $j \in [K]$ **do**

  $\widehat{\mathcal{M}}(\pi^{1:H,j}) \leftarrow$ `Construct`$(\pi^{1:H,j}, \{\widehat{r}_i^j\}_{i=1}^n, \{\widehat{\mathcal{C}}_h\}_{h\in[H+1]}, \{\widehat{\phi}_{h+1}\}_{h\in[H]}, \Gamma, \zeta_1, \zeta_2, \theta_1, \theta_2, \delta_1)$ // i.e., Algorithm 5

  $\pi^{\star,j} \leftarrow$ `VIACM`$(\widehat{\mathcal{M}}(\pi^{1:H,j}), \epsilon_e)$ // i.e., Algorithm 3

end for

$\pi^{\star,\widehat{j}} \leftarrow$ `PoS`$(\{\widehat{\mathcal{M}}(\pi^{1:H,j})\}_{j=1}^K, \{\pi^{\star,j}\}_{j=1}^K, \epsilon_e, N_2)$ // i.e., Algorithm 6

return $\pi^{\star,\widehat{j}}$

D. Full Versions of the Results

D.1. Hardness of finding team optimum and NE/CE/CCE

Throughout, we mainly consider the NE, CE, and CCE as our solution concepts. However, in Dec-POMDPs, a special class of POSGs with common rewards, a more common and favorable objective is the *team optimum*, where all agents jointly maximize the same expected return. In normal-form games, this team-optimal policy can be achieved by choosing the entry with the largest value in the payoff matrix and naturally extends to cooperative Markov games, when the model is known and the algorithm is centralized. However, in Dec-POMDPs, when partial observations appear, computing even approximate team optimal policies is `NEXP-complete` (Bernstein et al., 2002; Rabinovich et al., 2003), which means algorithms as in (Hansen et al., 2004) may take doubly exponential time in the worst case.

Then a natural question arises: if Dec-POMDPs have some favorable information-sharing structures, is computing the team-optimal policy still intractable? Unfortunately, even the agents share *all the information* without delays, which makes this Dec-POMDP a centralized POMDP, it is still `PSPACE-complete` to find the (team-)optimal solution (Papadimitriou & Tsitsiklis, 1987).

Recently, (Golowich et al., 2022b) considers *observable* POMDPs that rule out the ones with uninformative observations, for which computationally (quasi)-efficient algorithms can be developed. For POSGs including Dec-POMDPs, we could make a similar observability assumption, Assumption 2, on the joint observations, in the hope of obtaining computational (quasi)-efficiency.

This assumption only holds for the undercomplete setting where $S \leq O$. For the overcomplete setting, whether there also exists a computationally tractable algorithm is still open even for POMDPs. In fact, this observability assumption is equivalent (up to a factor of at most $\sqrt{O}$) to the ϵ-weakly revealing condition in (Liu et al., 2022b), under which there also exists *statistically* efficient algorithms. Directly adopting the main conclusion from (Golowich et al., 2022b), we can obtain the guarantee that with this full information sharing, there exists a quasi-polynomial algorithm computing the approximate team-optimal policy (see Proposition 4).

Given this simple positive result, one may wonder for Dec-POMDPs if we relax the strict requirement of *fully sharing*, but only with *partial information sharing* (but still under Assumption 2), is computing the team optimal policy still tractable? Unfortunately, we show that even under the strong sharing structure of only *one-step* delayed sharing, computing the team optimal policy is `NP-hard`. The formal proposition is introduced here.

Proposition 2. With 1-step delayed information-sharing structure and Assumption 2, computing the team optimal policy in Dec-POMDPs with $n = 2$ is `NP-hard`.

Proposition 2 implies that the standard observability assumption as in both single- (Golowich et al., 2022a) and multi-agent (Liu et al., 2022b) partially observable RL (i.e., Assumption 2) is not enough. Hence, instead of finding the overall team-optimum, hereafter we will mainly focus on finding the approximate *equilibrium* solutions (also known as *person-by-person* optimum in the team decision literature (Radner, 1962; Ho, 1980)). In particular, we focus on finding the NE in zero-sum or cooperative games, and CE/CCE in general-sum games, which are weaker notions than the team optimal solution in Dec-POMDPs.

Although the tractability of NE/CE/CCE in both zero-sum and general-sum normal-form games has been extensively studied, its formal tractability in POSGs has been less studied. Here by the following proposition, we will show that both Assumption 2 and some favorable information-sharing structure are necessary for NE/CE/CCE to be a computationally tractable solution concept even for zero-sum or cooperative POSGs, the proof of which is deferred to §E.1.

Proposition 3. For zero-sum or cooperative POSGs with only information sharing structures, or only Assumption 2, but not both, computing ϵ-NE/CE/CCE is `PSPACE-hard`.

This proposition shows that in order to seek a tractable algorithm even in zero-sum or cooperative POSGs, and even for the approximate and more relaxed solution concepts as CE/CCE, the condition of informative observations in Assumption 2 and the sharing of information are both necessary, in the sense that missing either one of them would make seeking approximate NE/CE/CCE computationally hard.

D.2. Planning

Here we state and prove our claims regarding planning in POSGs with the fully-sharing structure in §4.1.

Proposition 4. Let $\epsilon > 0$. Suppose the POSG $\mathcal{G}$ has a fully sharing structure and satisfies Assumption 2, then there is an algorithm that outputs an ϵ-suboptimal joint policy and has quasi-polynomial time complexity $H(AO)^{C\gamma^{-4}\log(\frac{SH}{\epsilon})}$ for some universal constant $C > 0$, where we recall γ is the constant appeared in Assumption 2.

To prove this, we first notice the following fact:

Fact 1. A Dec-POMDP with fully sharing structures can be solved by treating it as a (centralized) POMDP. The only difference is that in Dec-POMDPs, we care about policy $\pi_i \in \widetilde{\Pi}_i$. Meanwhile, a POMDP planning algorithm will provide a solution $\pi^\star \in \Pi^{\mathrm{gen}}$, where there could be a potential correlation among agents when taking actions. However, since there always exists deterministic solutions in POMDPs (Kaelbling et al., 1998) and as long as $\pi^\star$ is deterministic, it could be splitted into $\pi^\star = (\pi_1, \pi_2, \cdots, \pi_n)$ such that $\pi_i \in \widetilde{\Pi}_i$.

Hence, Proposition 4 comes from Fact 1 since the planning algorithm in (Golowich et al., 2022b) computes a deterministic policy, the Dec-POMDP is essentially a centralized POMDP, with joint observation and action space.

Below we state the full version of Theorem 2.

Theorem 6. Fix $\epsilon_r, \epsilon_z, \epsilon_e > 0$. Suppose there exists an (ϵ_r, ϵ_z)-expected-approximate common information model $\mathcal{M}$ for the POSG $\mathcal{G}$ that satisfies Assumption 3. Furthermore, if $\mathcal{M}$ is consistent with some given belief $\{\mathbb{P}_h^{\mathcal{M},c}(s_h, p_h|\widehat{c}_h)\}_{h\in[H]}$, then there exists a planning algorithm (Algorithm 1) outputting $\widehat{\pi}^\star$ such that NE-gap$(\widehat{\pi}^\star) \le 2H\epsilon_r + H^2\epsilon_z + H\epsilon_e$, if $\mathcal{G}$ is zero-sum or cooperative, and CE/CCE-gap$(\widehat{\pi}^\star) \le 2H\epsilon_r + H^2\epsilon_z + H\epsilon_e$ if $\mathcal{G}$ is general-sum, where the time complexity is $\max_{h\in[H]} \widehat{C}_h \cdot \mathrm{poly}(S, A, P_h, H, \frac{1}{\epsilon_e})$.

Now let us state the full version of Theorem 3.

Theorem 7. Fix $\epsilon > 0$. Suppose the POSG $\mathcal{G}$ satisfies Assumption 2. There exists a quasi-polynomial time algorithm computing ϵ-NE if $\mathcal{G}$ is zero-sum or cooperative and ϵ-CE/CCE if $\mathcal{G}$ is general-sum with the following information-sharing structures and time complexities, where we recall γ is the constant appeared in Assumption 2:

- **One-step delayed information sharing:** $(AO)^{C\gamma^{-4}\log\frac{SH}{\epsilon}}$ for some universal constant $C > 0$.
- **State controlled by one controller with asymmetric** $d = \mathrm{poly}(\log H)$**-step delayed sharing sharing:** $(AO)^{C(\gamma^{-4}\log\frac{SH}{\epsilon}+d)}$ for some constant $C > 0$.
- **Information sharing with one-directional-one-step delay:** $(AO)^{C\gamma^{-4}\log\frac{SH}{\epsilon}}$ for some universal constant $C > 0$.
- **Uncontrolled state process with** $d = \mathrm{poly}(\log H)$**-step delayed sharing:** $(AO)^{C(\gamma^{-4}\log\frac{SH}{\epsilon}+d)}$ for some universal constant $C > 0$.
- **Symmetric information game:** $(AO)^{C\gamma^{-4}\log\frac{SH}{\epsilon}}$ for some universal constant $C > 0$.

D.3. Learning

Here we state the full version of Theorem 4.

Theorem 8. Given compression function of common information, $\mathrm{Compress}_h : \mathcal{C}_h \to \widehat{\mathcal{C}}_h$ for $h \in [H]$, $\widehat{L}$ is as defined in Definition 7. Given H policies $\pi^{1:H}$, where $\pi^h \in \Pi^{\mathrm{gen}}$, $\pi^h_{h-\widehat{L}:h} = \mathrm{Unif}(\mathcal{A})$ for $h \in [H]$, and approximate reward functions $\widehat{r} = \{(\widehat{r}_{i,h})_{i=1}^n\}_{h=1}^H$, assume $\widetilde{\mathcal{M}}(\pi^{1:H})$ is an $(\epsilon_r(\pi^{1:H}, \widehat{r}), \epsilon_z(\pi^{1:H}))$-approximate common information model of $\mathcal{G}$ that satisfies Assumption 3. Fix some parameters $\delta_1, \theta_1, \theta_2, \zeta_1, \zeta_2 > 0$ for Algorithm 5, $\epsilon_e > 0$ for Algorithm 3, and $\phi > 0$, define the approximation error for estimating $\widetilde{\mathcal{M}}(\pi^{1:H})$ using samples under the policy $\pi^{1:H}$ as:

$$\epsilon_{apx}(\pi^{1:H}, \widehat{L}, \zeta_1, \zeta_2, \theta_1, \theta_2, \phi) = O\theta_1 + 2A\max_h P_h\frac{\zeta_2}{\zeta_1} + A\max_h P_h\theta_2 + \frac{A^{2\widehat{L}}O^{\widehat{L}}\zeta_1}{\phi} + \max_h \max_{\pi\in\Pi^{\mathrm{gen}}} \mathbb{1}[h > \widehat{L}]\cdot 2\cdot d^{\pi,\mathcal{G}}_{\mathcal{S},h-\widehat{L}}\Big(\mathcal{U}^{\mathcal{G}}_{\phi,h-\widehat{L}}(\pi^h)\Big),$$

where for any policy $\pi' \in \Pi^{\mathrm{gen}}, h \in [H], \phi > 0$, we define $d^{\pi',\mathcal{G}}_{\mathcal{S},h}(s) := \mathbb{P}_h^{\pi',\mathcal{G}}(s_h = s)$, $\mathcal{U}^{\mathcal{G}}_{\phi,h}(\pi') := \{s \in \mathcal{S} : d^{\pi',\mathcal{G}}_{\mathcal{S},h}(s) < \phi\}$, representing the under-explored states under the policy π'.

Then, Algorithm 5 can learn an model $\widehat{\mathcal{M}}(\pi^{1:H})$ with the sample complexity

$$N_0 = \max\left\{\frac{C(\max_h P_h + \log\frac{4H\max_h \widehat{C}_h}{\delta_1})}{\zeta_1\theta_1^2}, \frac{CA(O+\log\frac{4H\max_h(\widehat{C}_h P_h)A}{\delta_1})}{\zeta_2\theta_2^2}\right\}, \tag{D.1}$$

for some universal constant $C > 0$, such that with probability at least $1-\delta_1$, for any policy $\pi \in \Pi$, and $i \in [n]$:

$$\left|V_{i,1}^{\pi,\mathcal{G}}(\emptyset) - V_{i,1}^{\pi,\widehat{\mathcal{M}}(\pi^{1:H})}(\emptyset)\right| \le H\cdot\epsilon_r(\pi^{1:H},\widehat{r}) + \frac{H^2}{2}\epsilon_z(\pi^{1:H}) + \left(\frac{H^2}{2}+H\right)\epsilon_{apx}(\pi^{1:H},\widehat{L},\zeta_1,\zeta_2,\theta_1,\theta_2,\phi).$$

Furthermore, the policy output from the planning on $\widehat{\mathcal{M}}(\pi^{1:H})$ is an ϵ-NE if $\mathcal{G}$ zero-sum or cooperative and ϵ-CE/CCE if $\mathcal{G}$ is general-sum, where

$$\epsilon := H\epsilon_r(\pi^{1:H},\widehat{r}) + H^2\epsilon_z(\pi^{1:H}) + (H^2+H)\epsilon_{apx}(\pi^{1:H},\widehat{L},\zeta_1,\zeta_2,\theta_1,\theta_2,\phi) + H\epsilon_e.$$

Below we state the full version of Theorem 5.

Theorem 9. Fix $\epsilon,\delta > 0$. Suppose the POSG $\mathcal{G}$ satisfies Assumption 2. There exists a multi-agent RL algorithm (Algorithm 7) that learns an ϵ-NE if $\mathcal{G}$ is zero-sum or cooperative and ϵ-CE/CCE if $\mathcal{G}$ is general-sum with probability $1-\delta$, under the following information-sharing structures and corresponding sample and time complexities:

- **One-step delayed information sharing:** $(AO)^{C\gamma^{-4}\log\frac{SHO}{\gamma\epsilon}}\log\frac{1}{\delta}$ for some universal constant $C > 0$.
- **State controlled by one controller with asymmetric** $d = \mathrm{poly}(\log H)$**-step delayed sharing sharing:** $(AO)^{C(\gamma^{-4}\log\frac{SHO}{\gamma\epsilon}+d)}\log\frac{1}{\delta}$ for some constant $C > 0$.
- **Information sharing with one-directional-one-step delay:** $(AO)^{C\gamma^{-4}\log\frac{SHO}{\gamma\epsilon}}\log\frac{1}{\delta}$ for some universal constant $C > 0$.
- **Uncontrolled state process with** $d = \mathrm{poly}(\log H)$**-step delayed sharing:** $(AO)^{C(\gamma^{-4}\log\frac{SHO}{\gamma\epsilon}+d)}\log\frac{1}{\delta}$ for some universal constant $C > 0$.
- **Symmetric information game:** $(AO)^{C\gamma^{-4}\log\frac{SHO}{\gamma\epsilon}}\log\frac{1}{\delta}$ for some universal constant $C > 0$.

E. Technical Details and Omitted Proofs

E.1. Hardness of finding team optimum and NE/CE/CCE

To prove Proposition 2, we will firstly consider Dec-POMDPs with $H = 1$ and then connect the 1-step Dec-POMDP with Dec-POMDPs that have *1-step delayed sharing*. We will show the reduction from *Team Decision Problem* (Tsitsiklis & Athans, 1985):

Problem 1 (Team Decision Problem)**.** Given finite sets $\mathcal{Y}_1, \mathcal{Y}_2, \mathcal{U}_1, \mathcal{U}_2$, a rational probability function $p : \mathcal{Y}_1 \times \mathcal{Y}_2 \to \mathbb{Q}$ and an integer cost function $c : \mathcal{Y}_1 \times \mathcal{Y}_2 \times \mathcal{U}_1 \times \mathcal{U}_2 \to \mathbb{N}$, find decision rules $\gamma_i : \mathcal{Y}_i \to \mathcal{U}_i, i = 1, 2$, which minimize the expected cost:

$$J(\gamma_1, \gamma_2) = \sum_{y_1 \in \mathcal{Y}_1} \sum_{y_2 \in \mathcal{Y}_2} c(y_1, y_2, \gamma_1(y_1), \gamma_2(y_2)) p(y_1, y_2).$$

Proposition 5. Without any information sharing, computing jointly team optimal policies in Dec-POMDP with $H = 1$, $n = 2$ is `NP-hard`.

Proof. We can notice that the team decision problem is quite similar as our two-agent one-step Dec-POMDP. The only difference in Dec-POMDP is that the joint observations are sampled given the initial state, which is again sampled from μ_1. Now we will show how to reduce the team decision problem to a Dec-POMDP. For any team decision problem, we can construct the following Dec-POMDP:

- $\mathcal{A}_i = \mathcal{U}_i, i = 1, 2$
- $\mathcal{O}_i = \mathcal{Y}_i \cup (\mathcal{Y}_i \times \mathcal{U}_i), i = 1, 2$
- $\mathcal{S} = \mathcal{O}_1 \times \mathcal{O}_2$
- $\mathbb{O}(o_{1,h}, o_{2,h} | s_h) = 1$ if $s_h = (o_{1,h}, o_{2,h})$, else 0, for $h \in \{1, 2\}$.
- $r_2(o_{1,2}, o_{2,2}) = -c(y_1, y_2, a_1, a_2)$, where $o_{1,2} = (y_1, a_1)$ and $o_{2,2} = (y_2, a_2)$. We also define $r_1 = 0$.
- $\mu_1(s_1) = p(y_1, y_2)$, where $s_1 = (y_1, y_2)$
- $\mathbb{T}_1(s_2 | s_1, a_1, a_2) = 1$ if $s_2 = (s_1, a_1, a_2)$.

Based on the construction, computing the optimal policies $(\pi^\star_{1,1}, \pi^\star_{2,1})$ under the no-sharing structures in the reduced Dec-POMDP problem will give us the optimal policies $(\gamma^\star_1, \gamma^\star_2)$ in the original team decision problem. Concretely, $\gamma^\star_i(y_i) = \pi^\star_{i,1}(o_{i,1})$, where $o_{i,1} = y_i$. Given the `NP-hardness` of the team decision problem shown in (Tsitsiklis & Athans, 1985), solving the corresponding Dec-POMDP without information sharing is also `NP-hard`. □

This result immediately implies the hardness of Dec-POMDPs with 1-step delayed sharing structure:

Proposition 6. With 1-step delayed information-sharing structure, computing jointly team optimal policies in Dec-POMDPs with $n = 2$ is at least `NP-hard`.

Proof. Since there exists 1-step delay for the common information to be shared, when the Dec-POMDPs have only 1-step, there is no shared common information among agents. Therefore, based on Proposition 5, which concerns exactly such a case, computing joint optimal policies in Dec-POMDPs with $n = 2$ is also at least `NP-hard`. □

Finally, we are ready to prove Proposition 2.

Proof. Similar to the proof of Proposition 6, it suffices to show that the proposition holds for Dec-POMDPs, with $H = 1$ and without information sharing. Note in the proof of Proposition 5, the constructed Dec-POMDPs has the state space defined as the joint observation space (the Cartesian product of the individual observation space), the observation emission is actually a one-to-one mapping from state space to joint observation space. Correspondingly, $\mathbb{O}_h$ is indeed an identity matrix. Therefore, we have $\left\| \mathbb{O}_h^\top b - \mathbb{O}_h^\top b' \right\|_1 = \|b - b'\|_1$, for any $b, b' \in \Delta(\mathcal{S})$, verifying that $\gamma = 1$. □

Now let us restate and prove our hardness results regarding NE/CE/CCE in Proposition 3 as the following two propositions.

Proposition 7. For zero-sum or cooperative POSGs with any kind of information-sharing structure (including the fully-sharing structure), computing ϵ-NE/CE/CCE is `PSPACE-hard`.

Proof. The proof leverages the known results of the hardness of solving POMDPs. Given any instance of POMDPs, one could add a dummy player with only one dummy observation and one available action, which does not affect the transition, and use any desired information-sharing strategy. Since this dummy player only has one action and therefore it has only one policy. And the reward could be identical to the original player for cooperative games or the opposite of that for zero-sum games. Therefore, ϵ-NE/CE/CCE in this constructed POSGs with desired information-sharing strategy gives the ϵ-optimal policy in the original POMDP. Given the known `PSPACE-hardness` of POMDPs (Papadimitriou & Tsitsiklis, 1987; Lusena et al., 2001), we conclude our proof. □

Proposition 8. For zero-sum or cooperative POSGs satisfying Assumption 2 without information sharing, computing ϵ-NE/CE/CCE is `PSPACE-hard`.

Proof. Similar to the proof of Proposition 7, given any instance of POMDPs, we could add a dummy player with only one available action, and the observation of the dummy player is exactly the underlying state. Formally, given an instance of POMDP $\mathcal{P} = (\mathcal{S}^{\mathcal{P}}, \mathcal{A}^{\mathcal{P}}, \mathcal{O}^{\mathcal{P}}, \{\mathbb{O}_h^{\mathcal{P}}\}_{h\in[H+1]}, \{\mathbb{T}_h^{\mathcal{P}}\}_{h\in[H+1]}, r^{\mathcal{P}})$, we construct the POSG $\mathcal{G}$ as follows:

- $\mathcal{S} = \mathcal{S}^{\mathcal{P}}$.
- $\mathcal{A}_1 = \mathcal{A}^{\mathcal{P}}$, and $\mathcal{A}_2 = \{\emptyset\}$.
- $\mathcal{O}_1 = \mathcal{O}^{\mathcal{P}}$, and $\mathcal{O}_2 = \mathcal{S}^{\mathcal{P}}$.
- For any $h \in [H+1]$, $o_{1,h} \in \mathcal{O}_1$, $o_{2,h} \in \mathcal{O}_2$, $s_h \in \mathcal{S}$, it holds that

$$\mathbb{O}_h(o_{1,h}, o_{2,h}\,|\,s_h) = \begin{cases} \mathbb{O}_h^{\mathcal{P}}(o_{1,h}\,|\,s_h) & \text{if } o_{2,h} = s_h \\ 0 & \text{otherwise}, \end{cases}$$

- For any $h \in [H+1]$, $a_{1,h} \in \mathcal{A}_1$, $a_{2,h} \in \mathcal{A}_2$, $s_h, s_{h+1} \in \mathcal{S}$, it holds that $\mathbb{T}_h(s_{h+1}\,|\,s_h, a_{1,h}, a_{2,h}) = \mathbb{T}_h^{\mathcal{P}}(s_{h+1}\,|\,s_h, a_{1,h})$.
- For any $h \in [H+1]$, $o_{1,h} \in \mathcal{O}_1$, $o_{2,h} \in \mathcal{O}_2$, it holds that $r_{1,h}(o_{1,h}, o_{2,h}) = r_h^{\mathcal{P}}(o_{1,h})$, and $r_2 = r_1$ for cooperative games and $r_2 = -r_1$ for zero-sum games.

Now we are ready to verify that the joint observation emission satisfies Assumption 2 with $\gamma = 1$. Consider any $b, b' \in \Delta(\mathcal{S})$, denote $b - b' = (\delta_s)_{s\in\mathcal{S}}^\top$ as the column vector. For any $h \in [H+1]$, it holds that

$$\|\mathbb{O}_h^\top(b - b')\|_1 = \sum_{o_{1,h}, o_{2,h}} |\sum_{s\in\mathcal{S}} \mathbb{O}_h(o_{1,h}, o_{2,h}\,|\,s)\delta_s| = \sum_{o_{1,h}, o_{2,h}} |\mathbb{O}_h(o_{1,h}\,|\,o_{2,h})\delta_{o_{2,h}}| = \sum_{o_{2,h}} |\delta_{o_{2,h}}| = \|b - b'\|,$$

which verifies that $\gamma = 1$ for our constructed POSG. Computing ϵ-NE/CE/CCE in such a 1-observable POSG immediately gives us the ϵ-optimal policy in the original POMDP. Furthermore, note that $\gamma \le 1$ for any possible emission, therefore, the conclusion also holds for any γ-observable POSG, which proves our conclusion. □

Finally, we provide the proof for Lemma 1.

Proof. Fix any $h \in [H+1]$. If each player has perfect recall, then it holds that for any joint history $(a_1, o_2, \cdots, o_h) \in \mathcal{O}^{h-1} \times \mathcal{A}^{h-1}$, there exists some $c_h \in \mathcal{C}_h$ and $p_h \in \mathcal{P}_h$ such that $(c_h, p_h) = (a_1, o_2, \cdots, a_{h-1}, o_h)$. Therefore, we conclude that $\mathcal{O}^{h-1} \times \mathcal{A}^{h-1} \subseteq \mathcal{C}_h \times \mathcal{P}_h$, implying that $C_h P_h \ge (OA)^{h-1}$. □

E.2. Common information-based value iteration for POSGs

Similar to the value iteration algorithm in Markov games (Shapley, 1953), which solves a normal-form game at each step, we utilize a similar value iteration framework. However, it is not really possible to utilize the structure in Equation (2.4) to perform backward induction since $\{\mathbb{P}_h^{\pi_{1:h-1},\mathcal{G}}(s_h, p_h\,|\,c_h)\}_{h\in[H]}$ has dependency on the past policies $\pi_{1:h-1}$. Therefore, to compute $\pi_h^\star$, one not only needs to know $\pi_{h+1:H}^\star$ but also $\pi_{1:h-1}^\star$ because of the dependence of $\mathbb{P}_h^{\pi_{1:h-1},\mathcal{G}}$ on $\pi_{1:h-1}$. However, with Assumption 3, we can actually avoid the dependency on past policies and have

$$V_{i,h}^{\pi,\mathcal{G}}(c_h) = \mathbb{E}_{\{\omega_{j,h}\}_{j\in[n]}} \mathbb{E}_{s_h, p_h \sim \mathbb{P}_h^{\mathcal{G}}(\cdot,\cdot\,|\,c_h)} \mathbb{E}_{\{a_{j,h}\sim\pi_{j,h}(\cdot\,|\,\omega_{j,h}, c_h, p_{j,h})\}_{j\in[n]}} \mathbb{E}_{o_{h+1}\sim\mathbb{O}_{h+1}^\top \mathbb{T}_h(\cdot\,|\,s_h, a_h)} [r_{i,h+1}(o_{h+1}) + V_{i,h+1}^{\pi,\mathcal{G}}(c_{h+1})].$$

With Assumption 3, we are ready to present our Algorithm 1.

Now we will discuss the three equilibrium or best response (BR) subroutines we consider, where NE or NE-BR is used for zero-sum or cooperative games, and CE/CCE (or CE/CCE-BR) is used for general-sum games for computation tractability. To find efficient implementation for these subroutines, we need the following important properties on the prescription-value function.

Proposition 9. $Q_{i,h}^{\star,\mathcal{G}}(c_h, \gamma_{1,h}, \cdots, \gamma_{n,h})$ defined in Algorithm 1 is linear with respect to each $\gamma_{i,h}$. More specifically, we have:

$$\frac{\partial Q_{i,h}^{\star,\mathcal{G}}(c_h, \gamma_{1,h}, \cdots, \gamma_{n,h})}{\partial \gamma_{i,h}(a_{i,h}|p_{i,h})} = \sum_{s'_h, p'_{-i,h}} \sum_{a'_{-i,h}} \mathbb{P}_h^{\mathcal{G}}(s'_h, p_{i,h}, p'_{-i,h}|c_h)\gamma_{-i,h}(a'_{-i,h}|p'_{-i,h}) \sum_{o_{h+1}, s'_{h+1}} \mathbb{O}_{h+1}(o_{h+1}|s'_{h+1})\mathbb{T}_h(s'_{h+1}|s'_h, a_h)[r_{i,h+1}(o_{h+1}) + V_{i,h+1}^{\star,\mathcal{G}}(c_{h+1})]. \tag{E.1}$$

Proof. The partial derivative can be easily verified by algebraic manipulations and the definition of $Q_{i,h}^{\star,\mathcal{G}}$. From Equation (E.1), we could notice $\gamma_{i,h}$ does not appear on the RHS, which proves $Q_{i,h}^{\star,\mathcal{G}}(c_h, \gamma_{1,h}, \cdots, \gamma_{n,h})$ is linear with respect to $\gamma_{i,h}$. □

With such kind of linear structures, we are ready to introduce how to implement those oracles efficiently.

- The NE subroutine will give us the approximate NE $\gamma_{1,h}^\star, \cdots, \gamma_{n,h}^\star$ up to some error ϵ_e, which satisfies:

$$Q_{i,h}^{\star,\mathcal{G}}(c_h, \gamma_{i,h}^\star, \gamma_{-i,h}^\star) \geq \max_{\gamma_{i,h} \in \Delta(\mathcal{A}_i)^{P_{i,h}}} Q_{i,h}^{\star,\mathcal{G}}(c_h, \gamma_{i,h}, \gamma_{-i,h}^\star) - \epsilon_e, \forall i \in [n].$$

  This NE subroutine will be intractable for general-sum games even with only two players (Daskalakis et al., 2009; Chen et al., 2009). However, for cooperative games and zero-sum games, this NE subroutine can be implemented efficiently. At first look, this can be done by formulating it as a normal-form game, where each agent has the corresponding action space $\mathcal{A}_i^{P_{i,h}}$. However, this could not be tractable since the action space is indeed exponentially large. Fortunately, for cooperative games and two-player zero-sum games, we could utilize the linear (concave) structure, where $\gamma_{i,h}$ is a vector of dimension $A_i P_{i,h}$ to develop an efficient algorithm to compute ϵ_e-NE using standard no-external-regret or specifically gradient-play algorithms (Daskalakis et al., 2011; Zhang et al., 2021c; Leonardos et al., 2022; Ding et al., 2022; Mao et al., 2022), which will run in $\text{poly}(S, A, P_h, \frac{1}{\epsilon_e})$ time. To further illustrate how we avoid the dependence of $\mathcal{A}_i^{P_{i,h}}$, we refer to Figure 2. Similarly, the best response (BR) subroutine for NE, NE-BR subroutine will give us the approximate best response $\gamma_{i,h}^\star$ for agent i given $\{\gamma_{j,h}\}_{j\in[n]}$ up to some error ϵ_e, which satisfies:

$$Q_{i,h}^{\star,\mathcal{G}}(c_h, \gamma_{i,h}^\star, \gamma_{-i,h}) \geq \max_{\gamma'_{i,h} \in \Delta(\mathcal{A}_i)^{P_{i,h}}} Q_{i,h}^{\star,\mathcal{G}}(c_h, \gamma'_{i,h}, \gamma_{-i,h}) - \epsilon_e.$$

  Its implementation is straightforward by linear programming since $Q_{i,h}^{\star,\mathcal{G}}$ is linear with respect to each player's prescription.
- The CCE subroutine will give us the approximate CCE $\{\gamma_{1,h}^{\star,t}, \cdots, \gamma_{n,h}^{\star,t}\}_{t=1}^T$ up to some error ϵ_e, which satisfy:

$$\frac{1}{T}\sum_{t=1}^T Q_{i,h}^{\star,\mathcal{G}}(c_h, \gamma_{i,h}^{\star,t}, \gamma_{-i,h}^{\star,t}) \geq \max_{\gamma_{i,h} \in \Delta(\mathcal{A}_i)^{P_{i,h}}} \frac{1}{T}\sum_{t=1}^T Q_{i,h}^{\star,\mathcal{G}}(c_h, \gamma_{i,h}, \gamma_{-i,h}^{\star,t}) - \epsilon_e, \forall i \in [n].$$

  This subroutine can be implemented using standard no-external-regret learning algorithm as in (Gordon et al., 2008; Daskalakis et al., 2011) with $\text{poly}(S, A, P_h, \frac{1}{\epsilon_e})$ time.

  Similarly, the CCE-BR subroutine will give us the best response $\gamma_{i,h}^\star$ given $\{\gamma_{1,h}^t, \cdots, \gamma_{n,h}^t\}_{t=1}^T$ up to some error ϵ_e, which satisfies:

$$\frac{1}{T}\sum_{t=1}^T Q_{i,h}^{\star,\mathcal{G}}(c_h, \gamma_{i,h}^\star, \gamma_{-i,h}^t) \geq \max_{\gamma'_{i,h} \in \Delta(\mathcal{A}_i)^{P_{i,h}}} \frac{1}{T}\sum_{t=1}^T Q_{i,h}^{\star,\mathcal{G}}(c_h, \gamma'_{i,h}, \gamma_{-i,h}^t) - \epsilon_e.$$

  The implementation of CCE-BR is the same as CCE except that only the player i runs the no-external-regret algorithm and other players remain fixed. Once we get the sequence $\{\gamma_{i,h}^{\star,t}\}_{t=1}^T$ from the no-external-regret algorithm, we can take $\gamma_{i,h}^\star = \frac{1}{T}\sum_{t=1}^T \gamma_{i,h}^{\star,t}$ since $Q_{i,h}^{\star,\mathcal{G}}$ is linear with respect to each player's prescription.

- The CE subroutine will give us the approximate CE $\{\gamma_{1,h}^{\star,t}, \cdots, \gamma_{n,h}^{\star,t}\}_{t=1}^{T}$ up to some error ϵ_e, which satisfy:

$$\frac{1}{T}\sum_{t=1}^{T} Q_{i,h}^{\star,\mathcal{G}}(c_h, \gamma_{i,h}^{\star,t}, \gamma_{-i,h}^{\star,t}) \geq \max_{u_{i,h}} \frac{1}{T}\sum_{t=1}^{T} Q_{i,h}^{\star,\mathcal{G}}(c_h, u_{i,h} \diamond \gamma_{i,h}^{\star,t}, \gamma_{-i,h}^{\star,t}) - \epsilon_e, \forall i \in [n].$$

Here $u_{i,h} = \{u_{i,h,p_{i,h}}\}_{p_{i,h}}$ is the strategy modification, where $u_{i,h,p_{i,h}} : \mathcal{A}_i \to \mathcal{A}_i$ will modify the action $a_{i,h}$ to $u_{i,h,p_{i,h}}(a_{i,h})$ given the private information $p_{i,h}$. It is easy to see that the composition of $u_{i,h}$ with any prescription $\gamma_{i,h}$ is equivalent to $(u_{i,h} \diamond \gamma_{i,h})(a_{i,h}|p_{i,h}) := \sum_{u_{i,h,p_{i,h}}(a'_{i,h})=a_{i,h}} \gamma_{i,h}(a'_{i,h}|p_{i,h})$. One can verify that $u_{i,h} \diamond \gamma_{i,h} = U \cdot \gamma_{i,h}$, for some matrix $U \in \mathbb{R}^{A_i P_{i,h} \times A_i P_{i,h}}$ (in a block diagonal form). Therefore, the composition of $u_{i,h}$ and $\gamma_{i,h}$ is indeed a linear transformation. Now, as long as the function $Q_{i,h}^{\star}(c_h, \gamma_{1,h}, \cdots, \gamma_{n,h})$ is concave with respect to each $\gamma_{i,h}$, one can run the *no-linear-regret* algorithm as in (Gordon et al., 2008) in $\text{poly}(S, A, P_h, \frac{1}{\epsilon_e})$ time, such that the time-averaged policy will give us the approximate CE.
The CE-BR subroutine will give us the best strategy modification $u_{i,h}^{\star}$ given $\{\gamma_{1,h}^{t}, \cdots, \gamma_{n,h}^{t}\}_{t=1}^{T}$ up to some error ϵ_e, which satisfies:

$$\frac{1}{T}\sum_{t=1}^{T} Q_{i,h}^{\star,\mathcal{G}}(c_h, u_{i,h}^{\star} \diamond \gamma_{i,h}^{t}, \gamma_{-i,h}^{t}) \geq \max_{u_{i,h}} \frac{1}{T}\sum_{t=1}^{T} Q_{i,h}^{\star,\mathcal{G}}(c_h, u_{i,h} \diamond \gamma_{i,h}^{t}, \gamma_{-i,h}^{t}) - \epsilon_e.$$

For notational convenience, we shall slightly abuse the notation, writing $\gamma_{i,h}^{\star,t} := u_{i,h}^{\star} \diamond \gamma_{i,h}^{t}$ for any $t \in [T]$ and we assume our CE-BR subroutine returns $\{u_{i,h}^{\star} \diamond \gamma_{i,h}^{t}\}_{t\in[T]}$ instead of $u_{i,h}^{\star}$. Its implementation still follows that of CE except that only the agent i runs the *no-linear-regret* algorithm.

E.3. Near optimality of policies with approximate common information

To prove the main theorem, Theorem 2, we will first bound the sub-optimality at each step h through the following two lemmas.

Lemma 2. Fix the input $\mathcal{M}$ and $\epsilon_e > 0$ for Algorithm 3. For any $h \in [H+1]$, $c_h \in \mathcal{C}_h$, and $\pi_i \in \Pi_i$, for computing approximate NE/CCE, the output of Algorithm 3, $\widehat{\pi}^{\star}$ satisfies that

$$V_{i,h}^{\pi_i \times \widehat{\pi}_{-i}^{\star}, \mathcal{M}}(c_h) \leq V_{i,h}^{\widehat{\pi}^{\star}, \mathcal{M}}(c_h) + (H+1-h)\epsilon_e.$$

Proof. Obviously, the proposition holds for $h = H+1$. Note that π_i does not share the randomness with $\widehat{\pi}_{-i}^{\star}$. In other words, the following $\omega'_{i,h}$ is independent of $\omega_{-i,h}$. Then, we have that

$$\begin{aligned}
V_{i,h}^{\pi_i \times \widehat{\pi}_{-i}^{\star}, \mathcal{M}}(c_h) &= \mathbb{E}_{\omega'_{i,h}} \mathbb{E}_{\{\omega_{j,h}\}_{j\in[n]}} \mathbb{E}^{\mathcal{M}}[\widehat{r}_{i,h+1}(o_{h+1}) + V_{i,h+1}^{\pi_i \times \widehat{\pi}_{-i}^{\star}, \mathcal{M}}(c_{h+1}) \,|\, \widehat{c}_h, \{\pi_{i,h}(\cdot|\omega'_{i,h}, c_h, \cdot), \widehat{\pi}_{-i,h}^{\star}(\cdot|\omega_{-i,h}, \widehat{c}_h, \cdot)\}] \\
&\leq \mathbb{E}_{\omega'_{i,h}} \mathbb{E}_{\{\omega_{j,h}\}_{j\in[n]}} \mathbb{E}^{\mathcal{M}}[\widehat{r}_{i,h+1}(o_{h+1}) + V_{i,h+1}^{\widehat{\pi}^{\star}, \mathcal{M}}(c_{h+1}) \,|\, \widehat{c}_h, \{\pi_{i,h}(\cdot|\omega'_{i,h}, c_h, \cdot), \widehat{\pi}_{-i,h}^{\star}(\cdot|\omega_{-i,h}, \widehat{c}_h, \cdot)\}] + (H-h)\epsilon_e \qquad \text{(E.2)} \\
&= \mathbb{E}_{\omega'_{i,h}} \mathbb{E}_{\{\omega_{j,h}\}_{j\in[n]}} Q_{i,h}^{\widehat{\pi}_i^{\star} \times \widehat{\pi}_{-i}^{\star}, \mathcal{M}}(c_h, \pi_{i,h}(\cdot|\omega'_{i,h}, c_h, \cdot), \widehat{\pi}_{-i,h}^{\star}(\cdot|\omega_{-i,h}, \widehat{c}_h, \cdot)) + (H-h)\epsilon_e \\
&\leq \mathbb{E}_{\omega'_{i,h}} \mathbb{E}_{\{\omega_{j,h}\}_{j\in[n]}} Q_{i,h}^{\widehat{\pi}_i^{\star} \times \widehat{\pi}_{-i}^{\star}, \mathcal{M}}(c_h, \widehat{\pi}_{i,h}^{\star}(\cdot|\omega_{i,h}, c_h, \cdot), \widehat{\pi}_{-i,h}^{\star}(\cdot|\omega_{-i,h}, \widehat{c}_h, \cdot)) + (H-h+1)\epsilon_e \qquad \text{(E.3)} \\
&= V_{i,h}^{\widehat{\pi}^{\star}, \mathcal{M}}(c_h) + (H-h+1)\epsilon_e.
\end{aligned}$$

Equation (E.2) comes from the inductive hypothesis. Equation (E.3) holds since $V_{i,h+1}^{\widehat{\pi}^{\star}, \mathcal{M}}(c_{h+1}) = V_{i,h+1}^{\widehat{\pi}^{\star}, \mathcal{M}}(\widehat{c}_{h+1})$ and $\widehat{\pi}_h^{\star}(\cdot \,|\, \cdot, \widehat{c}_h, \cdot)$ is an ϵ_e-NE/CCE. □

Lemma 3. For any $h \in [H]$, $c_h \in \mathcal{C}_h$, and $\phi_i \in \Phi_i$, for computing approximate CE, the output of Algorithm 3, $\widehat{\pi}^{\star}$ satisfies that

$$V_{i,h}^{(\phi_i \diamond \widehat{\pi}_i^{\star}) \odot \widehat{\pi}_{-i}^{\star}, \mathcal{M}}(c_h) \leq V_{i,h}^{\widehat{\pi}^{\star}, \mathcal{M}}(c_h) + (H-h+1)\epsilon_e.$$

Proof. Obviously, the proposition holds for $h = H+1$. Moreover, we have

$$
\begin{aligned}
&V_{i,h}^{(\phi_i \diamond \widehat{\pi}_i^\star) \odot \widehat{\pi}_{-i}^\star, \mathcal{M}}(c_h) = \mathbb{E}_{\{\omega_{j,h}\}_{j\in[n]}} \mathbb{E}^{\mathcal{M}}[\widehat{r}_{i,h+1}(o_{h+1}) + V_{i,h+1}^{(\phi_i \diamond \widehat{\pi}_i^\star) \odot \widehat{\pi}_{-i}^\star, \mathcal{M}}(c_{h+1}) \mid \widehat{c}_h, \{\phi_{i,h,c_h} \diamond \widehat{\pi}_{i,h}^\star(\cdot \mid \omega_{i,h}, \widehat{c}_h, \cdot), \widehat{\pi}_{-i,h}^\star(\cdot \mid \omega_{-i,h}, \widehat{c}_h, \cdot)\}] \\
&\le \mathbb{E}_{\{\omega_{j,h}\}_{j\in[n]}} \mathbb{E}^{\mathcal{M}}[\widehat{r}_{i,h+1}(o_{h+1}) + V_{i,h+1}^{\widehat{\pi}^\star, \mathcal{M}}(c_{h+1}) \mid \widehat{c}_h, \{\phi_{i,h,c_h} \diamond \widehat{\pi}_{i,h}^\star(\cdot \mid \omega_{i,h}, \widehat{c}_h, \cdot), \widehat{\pi}_{-i,h}^\star(\cdot \mid \omega_{-i,h}, \widehat{c}_h, \cdot)] + (H-h)\epsilon_e \qquad \text{(E.4)} \\
&\le \mathbb{E}_{\{\omega_{j,h}\}_{j\in[n]}} \mathbb{E}^{\mathcal{M}}[\widehat{r}_{i,h+1}(o_{h+1}) + V_{i,h+1}^{\widehat{\pi}^\star, \mathcal{M}}(c_{h+1}) \mid \widehat{c}_h, \{\widehat{\pi}_{i,h}^\star(\cdot \mid \omega_{i,h}, \widehat{c}_h, \cdot), \widehat{\pi}_{-i,h}^\star(\cdot \mid \omega_{-i,h}, \widehat{c}_h, \cdot)\}] + (H-h)\epsilon_e \qquad \text{(E.5)} \\
&= V_{i,h}^{\widehat{\pi}^\star, \mathcal{M}}(c_h) + (H-h+1)\epsilon_e.
\end{aligned}
$$

Equation (E.4) comes from the inductive hypothesis. Equation (E.5) holds since $V_{i,h+1}^{\widehat{\pi}^\star, \mathcal{M}}(c_{h+1}) = V_{i,h+1}^{\widehat{\pi}^\star, \mathcal{M}}(\widehat{c}_{h+1})$ and $\widehat{\pi}_h^\star(\cdot \mid \cdot, \widehat{c}_h, \cdot)$ is an ϵ_e-CE. □

Now we need the following lemma, showing the difference between the approximate value functions and true value functions under the same set of policies.

Lemma 4. For any given policy $\pi' \in \Pi^{\text{gen}}$, $\pi \in \Pi$, and $h \in [H+1]$, we have

$$
\mathbb{E}_{a_{1:h-1}, o_{2:h} \sim \pi'}[|V_{i,h}^{\pi, \mathcal{G}}(c_h) - V_{i,h}^{\pi, \mathcal{M}}(c_h)|] \le (H-h+1)\epsilon_r + \frac{(H-h+1)(H-h)}{2}\epsilon_z.
$$

Proof. Obviously, the proposition holds for $h = H+1$. Furthermore, we have

$$
\begin{aligned}
&\mathbb{E}_{a_{1:h-1}, o_{2:h} \sim \pi'}[|V_{i,h}^{\pi, \mathcal{G}}(c_h) - V_{i,h}^{\pi, \mathcal{M}}(c_h)|] \\
&\le \mathbb{E}_{a_{1:h-1}, o_{2:h} \sim \pi'}[|\mathbb{E}_{\{\omega_{j,h}\}_{j\in[n]}} \mathbb{E}^{\mathcal{G}}[r_{i,h+1}(o_{h+1}) \mid c_h, \{\pi_{j,h}(\cdot \mid \omega_{j,h}, c_h, \cdot)\}_{j=1}^n] - \mathbb{E}_{\{\omega_{j,h}\}_{j\in[n]}} \mathbb{E}^{\mathcal{M}}[\widehat{r}_{i,h+1}(o_{h+1}) \mid \widehat{c}_h, \{\pi_{j,h}(\cdot \mid \omega_{j,h}, c_h, \cdot)\}_{j=1}^n]|] \\
&\quad + \mathbb{E}_{a_{1:h-1}, o_{2:h} \sim \pi'}[|\mathbb{E}_{\{\omega_{j,h}\}_{j\in[n]}} \mathbb{E}_{z_{h+1} \sim \mathbb{P}_h^{\mathcal{G}}(\cdot \mid c_h, \{\pi_{j,h}(\cdot \mid \omega_{j,h}, c_h, \cdot)\}_{j=1}^n)}[V_{i,h+1}^{\pi, \mathcal{G}}(\{c_h, z_{h+1}\})] \\
&\quad - \mathbb{E}_{\{\omega_{j,h}\}_{j\in[n]}} \mathbb{E}_{z_{h+1} \sim \mathbb{P}_h^{\mathcal{M}, z}(\cdot \mid \widehat{c}_h, \{\pi_{j,h}(\cdot \mid \omega_{j,h}, c_h, \cdot)\}_{j=1}^n)}[V_{i,h+1}^{\pi, \mathcal{M}}(\{c_h, z_{h+1}\})]|] \\
&\le \epsilon_r + (H-h)\mathbb{E}_{a_{1:h-1}, o_{2:h} \sim \pi'} \mathbb{E}_{\{\omega_{j,h}\}_{j\in[n]}} \|\mathbb{P}_h^{\mathcal{G}}(\cdot \mid c_h, \{\pi_{j,h}(\cdot \mid \omega_{j,h}, c_h, \cdot)\}_{j=1}^n) - \mathbb{P}_h^{\mathcal{M}, z}(\cdot \mid \widehat{c}_h, \{\pi_{j,h}(\cdot \mid \omega_{j,h}, c_h, \cdot)\}_{j=1}^n)\|_1 \\
&\quad + \mathbb{E}_{a_{1:h}, o_{2:h+1} \sim (\pi'_{1:h-1}, \pi_{h:H})}[|V_{i,h+1}^{\pi, \mathcal{M}}(c_{h+1}) - V_{i,h+1}^{\pi, \mathcal{G}}(c_{h+1})|] \\
&\le \epsilon_r + (H-h)\epsilon_z + (H-h)\epsilon_r + \frac{(H-h-1)(H-h)}{2}\epsilon_z \\
&\le (H-h+1)\epsilon_r + \frac{(H-h)(H-h+1)}{2}\epsilon_z,
\end{aligned}
$$

which completes the proof. □

Finally, we are ready to prove our main theorem, Theorem 2.

Proof. For computing NE/CCE, we define

$$
\pi_i^\star \in \arg\max_{\pi_i \in \Pi_i} V_{i,1}^{\pi_i \times \widehat{\pi}_{-i}^\star, \mathcal{M}}(\emptyset).
$$

Now note that

$$
\begin{aligned}
&\mathbb{E}_{a_{1:h-1}, o_{2:h} \sim \pi'}[V_{i,h}^{\pi_i^\star \times \widehat{\pi}_{-i}^\star, \mathcal{G}}(c_h) - V_{i,h}^{\widehat{\pi}^\star, \mathcal{G}}(c_h)] \\
&= \mathbb{E}_{a_{1:h-1}, o_{2:h} \sim \pi'}\left[\left(V_{i,h}^{\pi_i^\star \times \widehat{\pi}_{-i}^\star, \mathcal{G}}(c_h) - V_{i,h}^{\widehat{\pi}^\star, \mathcal{M}}(c_h)\right) + \left(V_{i,h}^{\widehat{\pi}^\star, \mathcal{M}}(c_h) - V_{i,h}^{\widehat{\pi}^\star, \mathcal{G}}(c_h)\right)\right] \\
&\le \mathbb{E}_{a_{1:h-1}, o_{2:h} \sim \pi'}\left[\left(V_{i,h}^{\pi_i^\star \times \widehat{\pi}_{-i}^\star, \mathcal{G}}(c_h) - V_{i,h}^{\pi_i^\star \times \widehat{\pi}_{-i}^\star, \mathcal{M}}(c_h)\right) + \left(V_{i,h}^{\widehat{\pi}^\star, \mathcal{M}}(c_h) - V_{i,h}^{\widehat{\pi}^\star, \mathcal{G}}(c_h)\right)\right] + (H+1-h)\epsilon_e \\
&\le 2(H-h+1)\epsilon_r + (H-h)(H-h+1)\epsilon_z + (H-h+1)\epsilon_e.
\end{aligned}
$$

Let $h = 1$, and note that $c_1 = \emptyset$, we get

$$V_{i,1}^{\pi_i^\star \times \widehat{\pi}_{-i}^\star, \mathcal{G}}(\emptyset) - V_{i,1}^{\widehat{\pi}^\star, \mathcal{G}}(\emptyset) \le 2H\epsilon_r + H^2\epsilon_z + H\epsilon_e.$$

By the definition of $\pi_i^\star$, we conclude

$$\text{NE/CCE-gap}(\widehat{\pi}^\star) \le 2H\epsilon_r + H^2\epsilon_z + H\epsilon_e.$$

For computing CE, define

$$\phi_i^\star \in \arg\max_{\phi_i} V_{i,1}^{(\phi_i \diamond \widehat{\pi}_i^\star) \odot \widehat{\pi}_{-i}^\star, \mathcal{M}}(\emptyset).$$

Now note that

$$\begin{aligned}
&\mathbb{E}_{a_{1:h-1}, o_{2:h} \sim \pi'}[V_{i,h}^{(\phi_i^\star \diamond \widehat{\pi}_i^\star) \odot \widehat{\pi}_{-i}^\star, \mathcal{G}}(c_h) - V_{i,h}^{\widehat{\pi}^\star, \mathcal{G}}(c_h)] \\
&= \mathbb{E}_{a_{1:h-1}, o_{2:h} \sim \pi'}\left[\left(V_{i,h}^{(\phi_i^\star \diamond \widehat{\pi}_i^\star) \odot \widehat{\pi}_{-i}^\star, \mathcal{G}}(c_h) - V_{i,h}^{\widehat{\pi}^\star, \mathcal{M}}(c_h)\right) + \left(V_{i,h}^{\widehat{\pi}^\star, \mathcal{M}}(c_h) - V_{i,h}^{\widehat{\pi}^\star, \mathcal{G}}(c_h)\right)\right] \\
&\le \mathbb{E}_{a_{1:h-1}, o_{2:h} \sim \pi'}\left[\left(V_{i,h}^{(\phi_i^\star \diamond \widehat{\pi}_i^\star) \odot \widehat{\pi}_{-i}^\star, \mathcal{G}}(c_h) - V_{i,h}^{(\phi_i^\star \diamond \widehat{\pi}_i^\star) \odot \widehat{\pi}_{-i}^\star, \mathcal{M}}(c_h)\right) + \left(V_{i,h}^{\widehat{\pi}^\star, \mathcal{M}}(c_h) - V_{i,h}^{\widehat{\pi}^\star, \mathcal{G}}(c_h)\right)\right] + (H + 1 - h)\epsilon_e \\
&\le 2(H - h + 1)\epsilon_r + (H - h)(H - h + 1)\epsilon_z + (H - h + 1)\epsilon_e.
\end{aligned}$$

Let $h = 1$, and note that $c_1 = \emptyset$, we get

$$V_{i,1}^{(\phi_i^\star \diamond \widehat{\pi}_i^\star) \odot \widehat{\pi}_{-i}^\star, \mathcal{G}}(\emptyset) - V_{i,1}^{\widehat{\pi}^\star, \mathcal{G}}(\emptyset) \le 2H\epsilon_r + H^2\epsilon_z + H\epsilon_e.$$

By the definition of $\phi_i^\star$, we conclude

$$\text{CE-gap}(\widehat{\pi}^\star) \le 2H\epsilon_r + H^2\epsilon_z + H\epsilon_e.$$

The last step is the analysis of the computation complexity. A major difference with the exact common information setting is that it is unclear whether there exists efficient NE/CE/CCE subroutines at each step h. However, if $\mathcal{M}$ is consistent with some approximate belief $\{\mathbb{P}_h^{\mathcal{M},c}(s_h, p_h | \widehat{c}_h)\}_{h \in [H]}$, through exactly the same argument as in Proposition 9 with $\mathbb{P}_h^{\mathcal{G}}(s_h, p_h | c_h)$ replaced by $\mathbb{P}_h^{\mathcal{M},c}(s_h, p_h | \widehat{c}_h)$, we conclude the NE subroutine for zero-sum or cooperative games and CE/CCE subroutine for general-sum games can be implemented efficiently in the computation complexity of $\text{poly}(S, A, P_h, \frac{1}{\epsilon_e})$. Now computation complexity of the Algorithm 3 is $H \max_h \widehat{C}_h \text{poly}(S, A, P_h, \frac{1}{\epsilon_e})$, where $\widehat{C}_h$ comes from the loop at each step h. □

E.4. Approximate common information with finite memory

Theorem 2 provides a structural result for the optimality of NE/CE/CCE policy computed with approximate common information in the underlying POSG when the approximate common information satisfies the condition in Definition 5. However, it is not clear how to construct such approximate common information and how high the induced computational complexity is. Here we will show when the joint observation is informative enough, specifically satisfying Assumption 2, we could simply use truncation to compress the common information and the corresponding most recent L steps of history is indeed a kind of approximate common information. Here we need the following result showing that most recent history is enough to predict the latent state of the POSG. Here we shall use a slightly stronger argument than (Golowich et al., 2022b), since we need to allow the policy $\pi' \in \Pi^{\text{gen}}$ to be stochastic while in POMDPs, deterministic policies are enough for optimal solutions. The proof goes quite similar to that in (Golowich et al., 2022b). Firstly, we shall need the following important lemmas.

Lemma 5 (Lemma 4.9 in (Golowich et al., 2022b))**.** Suppose the POSG satisfies Assumption 2, $b, b' \in \Delta(\mathcal{S})$ with $b \ll b'$, and fix any $h \in [H]$. Then

$$\mathbb{E}_{y \sim \mathbb{O}_h^\top b}\left[\sqrt{\exp\left(\frac{D_2(B_h(b;y) \| B_h(b';y))}{4}\right) - 1}\right] \le \left(1 - \gamma^4/2^{40}\right) \cdot \sqrt{\exp\left(\frac{D_2(b \| b')}{4}\right) - 1}.$$

This lemma states that once the emission $\mathbb{O}_h$ satisfies the condition in Assumption 2, the Bayes operator B_h is a contraction in expectation. Since the individual emission $\mathbb{O}_{i,h}$ does not necessarily satisfy Assumption 2, the individual Bayes operator $B_{i,h}$ follows a weaker proposition. We first state a more generalized lemma as follows.

Lemma 6. Given two finite domains X, Y, and the conditional probability $q(y|x)$ for $x \in X, y \in Y$. Define the posterior update $F^q(P;y) : \Delta(X) \to \Delta(X)$ for $P \in \Delta(X), y \in Y$ as

$$F^q(P;y)(x) = \frac{P(x)q(y|x)}{\sum_{x' \in X} P(x')q(y|x')}. \tag{E.6}$$

Then for any $\delta_1, \delta_2 \in \Delta(X)$ such that $\delta_1 \ll \delta_2$, it holds that

$$\mathbb{E}_{x\sim\delta_1, y\in q(\cdot|x)} \sqrt{\exp\left(\frac{D_2(F^q(\delta_1;y)\|F^q(\delta_2;y))}{4}\right) - 1} \le \sqrt{\exp\left(\frac{D_2(\delta_1\|\delta_2)}{4}\right) - 1}.$$

Proof. This is a direct consequence from the proof of Lemma 4.9 in (Golowich et al., 2022b) by allowing $\gamma = 0$ since here we do not assume any observability on q. □

Corollary 1. Suppose $b, b' \in \Delta(\mathcal{S})$ with $b \ll b'$, and fix any $h \in [H], i \in [n]$. Then

$$\mathbb{E}_{y\sim\mathbb{O}_{i,h}^\top b}\left[\sqrt{\exp\left(\frac{D_2\left(B_{i,h}(b;y)\|B_{i,h}(b';y)\right)}{4}\right) - 1}\right] \le \sqrt{\exp\left(\frac{D_2\left(b\|b'\right)}{4}\right) - 1}.$$

Lemma 7 (Lemma 4.8 in (Golowich et al., 2022b))**.** Consider probability distributions P, Q. Then

$$\|P - Q\|_1 \le 4 \cdot \sqrt{\exp\left(D_2(P\|Q)/4\right) - 1}.$$

Theorem 10 (Adapted from Theorem 4.7 in (Golowich et al., 2022b))**.** There is a constant $C \ge 1$ so that the following holds. Suppose that the POSG satisfies Assumption 2 with parameter γ. Let $\epsilon \ge 0$. Fix a policy $\pi' \in \Pi^{\mathrm{gen}}$ and indices $1 \le h - L < h - 1 \le H$. If $L \ge C\gamma^{-4}\log(\frac{S}{\epsilon})$, then the following set of propositions hold

$$\mathbb{E}^{\mathcal{G}}_{a_{1:h-1}, o_{2:h}\sim\pi'} \|\boldsymbol{b}_h(a_{1:h-1}, o_{2:h-1}) - \boldsymbol{b}'_h(a_{h-L:h-1}, o_{h-L+1:h-1})\|_1 \le \epsilon. \tag{E.7}$$

$$\mathbb{E}^{\mathcal{G}}_{a_{1:h-1}, o_{2:h}\sim\pi'} \|\boldsymbol{b}_h(a_{1:h-1}, o_{2:h}) - \boldsymbol{b}'_h(a_{h-L:h-1}, o_{h-L+1:h})\|_1 \le \epsilon. \tag{E.8}$$

$$\mathbb{E}^{\mathcal{G}}_{a_{1:h-1}, o_{2:h}\sim\pi'} \|\boldsymbol{b}_h(a_{1:h-1}, o_{2:h-1}, o_{1,h}) - \boldsymbol{b}'_h(a_{h-L:h-1}, o_{h-L+1:h-1}, o_{1,h})\|_1 \le \epsilon. \tag{E.9}$$

Furthermore, for any finite domain Y, conditional probability $q(y|s)$ and the posterior update operator $F^q : \Delta(\mathcal{S}) \to \Delta(\mathcal{S})$ as defined in Lemma 6, it holds that

$$\mathbb{E}^{\mathcal{G}}_{a_{1:h-1}, o_{2:h}\sim\pi'} \mathbb{E}_{y\sim q\cdot\boldsymbol{b}_h(a_{1:h-1}, o_{2:h})} \|F^q(\boldsymbol{b}_h(a_{1:h-1}, o_{2:h}); y) - F^q(\boldsymbol{b}'_h(a_{h-L:h-1}, o_{h-L+1:h}); y)\|_1 \le \epsilon. \tag{E.10}$$

Proof. Let us prove (E.7) first. Note that if $h - L \le 1$, then we have $\boldsymbol{b}_h(a_{1:h-1}, o_{2:h-1}) = \boldsymbol{b}'_h(a_{h-L:h-1}, o_{h-L+1:h-1})$. The proposition trivially holds. Now let us consider $h > L + 1$. Fix some history $(a_{1:h-L-1}, o_{2:h-L-1})$. We condition on this history throughout the proof. For $0 \le t \le L$, define the random variables

$$\begin{aligned}
b_{h-L+t} &= \boldsymbol{b}_{h-L+t}(a_{1:h-L+t-1}, o_{2:h-L+t-1}), \\
b'_{h-L+t} &= \boldsymbol{b}'_{h-L+t}(a_{h-L:h-L+t-1}, o_{h-L+1:h-L+t-1}), \\
Y_t &= \sqrt{\exp\left(\frac{D_2\left(b_{h-L+t}\|b'_{h-L+t}\right)}{4}\right) - 1}.
\end{aligned}$$

Then $D_2(b_{h-L}\|b'_{h-L}) = \log \mathbb{E}_{x\sim b_h} \frac{b_h(x)}{b'_h(x)} \le \log(S)$ since $b_{h-L} = \boldsymbol{b}'_{h-L}(\emptyset) = \mathrm{Unif}(\mathcal{S})$, so we have $Y_0 \le \sqrt{\exp(D_2(b_{h-L}\|b'_{h-L}))} \le S$. Moreover, for any $0 \le t \le L-1$, we have

$$
\begin{aligned}
&\mathbb{E}_{a_{h-L:h-L+t}, o_{h-L+1:h-L+t}\sim\pi'} Y_{t+1} = \mathbb{E}_{a_{h-L:h-L+t-1}, o_{h-L+1:h-L+t}\sim\pi'} \mathbb{E}_{a_{h-L+t}\sim\pi'(\cdot|a_{1:h-L+t-1}, o_{2:h-L+t})} \\
&\sqrt{\exp\left(\frac{D_2(\mathbb{T}_{h-L+t}(a_{h-L+t})\cdot B_{h-L+t}(b_{h-L+t}; o_{h-L+t})\|\mathbb{T}_{h-L+t}(a_{h-L+t})\cdot B_{h-L+t}(b'_{h-L+t}; o_{h-L+t}))}{4}\right)-1} \\
&\le \mathbb{E}_{a_{h-L:h-L+t-1}, o_{h-L+1:h-L+t-1}\sim\pi'} \mathbb{E}_{o_{h-L+t}\sim\mathbb{O}^\top_{h-L+t} b_{h-L+t}} \sqrt{\exp\left(\frac{D_2(B_{h-L+t}(b_{h-L+t}; o_{h-L+t})\|B_{h-L+t}(b'_{h-L+t}; o_{h-L+t}))}{4}\right)-1} \\
&\le \left(1-\frac{\gamma^4}{2^{40}}\right) \mathbb{E}_{a_{h-L:h-L+t-1}, o_{h-L+1:h-L+t-1}\sim\pi'} Y_t,
\end{aligned}
$$

where the second last step comes from the data processing inequality and the last step comes from Lemma 5. By induction and the choice of L, we have that

$$
\mathbb{E}_{o_{h-L:h-1}, a_{h-L:h-1}\sim\pi'} \sqrt{\exp\left(\frac{D_2\left(b_h\|b'_h\right)}{4}\right)-1} \le \left(1-\frac{\gamma^4}{2^{40}}\right)^L S \le \frac{\epsilon}{4}. \tag{E.11}
$$

It follows from Lemma 7 that

$$
\mathbb{E}_{a_{h-L:h-1}, o_{h-L+1:h-1}\sim\pi'} \|b_h - b'_h\|_1 \le \epsilon.
$$

Taking expectation over the history $(o_{h-L:h-1}, a_{h-L:h-1})$ completes the proof of (E.7). Finally, (E.8) follows from (E.11) and Lemma 5. (E.9) follows from (E.11) and Corollary 1. The (E.10) follows from (E.11) and Lemma 6. □

Before instantiating our information structure with particular cases, for convenience of our proof, we firstly identify a more sufficient condition for our Definition 5.

Lemma 8. Given any belief $\{\mathbb{P}_h^{\mathcal{M},c}(s_h, p_h|\widehat{c}_h)\}_{h\in[H]}$, assume $\mathcal{M}$ is consistent with $\{\mathbb{P}_h^{\mathcal{M},c}(s_h, p_h|\widehat{c}_h)\}_{h\in[H]}$. Then it holds that for any $h \in [H], c_h \in \mathcal{C}_h, \gamma_h \in \Gamma_h$:

$$
\|\mathbb{P}_h^{\mathcal{G}}(\cdot|c_h, \gamma_h) - \mathbb{P}_h^{\mathcal{M},z}(\cdot|\widehat{c}_h, \gamma_h)\|_1 \le \|\mathbb{P}_h^{\mathcal{G}}(\cdot,\cdot|c_h) - \mathbb{P}_h^{\mathcal{M},c}(\cdot,\cdot|\widehat{c}_h)\|_1, \tag{E.12}
$$

$$
|\mathbb{E}^{\mathcal{G}}[\widehat{r}_{i,h+1}(o_{h+1}) \mid c_h, \gamma_h] - \mathbb{E}^{\mathcal{M}}[\widehat{r}_{i,h+1}(o_{h+1}) \mid \widehat{c}_h, \gamma_h]| \le \|\mathbb{P}_h^{\mathcal{G}}(\cdot,\cdot|c_h) - \mathbb{P}_h^{\mathcal{M},c}(\cdot,\cdot|\widehat{c}_h)\|_1. \tag{E.13}
$$

Proof. Note that

$$
|\mathbb{E}^{\mathcal{G}}[\widehat{r}_{i,h+1}(o_{h+1}) \mid c_h, \gamma_h] - \mathbb{E}^{\mathcal{M}}[\widehat{r}_{i,h+1}(o_{h+1}) \mid \widehat{c}_h, \gamma_h]| \le \sum_{o_{h+1}} |\mathbb{P}_h^{\mathcal{G}}(o_{h+1}|c_h, \gamma_h) - \mathbb{P}_h^{\mathcal{M},o}(o_{h+1}|\widehat{c}_h, \gamma_h)|.
$$

Therefore, it suffices to bound $\sum_{o_{h+1}} |\mathbb{P}_h^{\mathcal{G}}(o_{h+1}|c_h, \gamma_h) - \mathbb{P}_h^{\mathcal{M},o}(o_{h+1}|\widehat{c}_h, \gamma_h)|$ for (E.13). Now, note that for any $c_h \in \mathcal{C}_h, \gamma_h \in \Gamma_h$:

$$
\begin{aligned}
&\sum_{s_h, p_h, a_h, s_{h+1}, o_{h+1}} |\mathbb{P}_h^{\mathcal{G}}(s_h, s_{h+1}, p_h, a_h, o_{h+1} \mid c_h, \gamma_h) - \mathbb{P}_h^{\mathcal{M}}(s_h, s_{h+1}, p_h, a_h, o_{h+1}|\widehat{c}_h, \gamma_h)| \\
&= \sum_{s_h, p_h, a_h, s_{h+1}, o_{h+1}} |\mathbb{P}_h^{\mathcal{G}}(s_h, p_h \mid c_h) \Pi_{j=1}^n \gamma_{j,h}(a_{j,h} \mid p_{j,h}) \mathbb{T}_h(s_{h+1}|s_h, a_h) \mathbb{O}_{h+1}(o_{h+1}|s_{h+1}) \\
&\quad - \mathbb{P}_h^{\mathcal{M},c}(s_h, p_h|\widehat{c}_h) \Pi_{j=1}^n \gamma_{j,h}(a_{j,h} \mid p_{j,h}) \mathbb{T}_h(s_{h+1}|s_h, a_h) \mathbb{O}_{h+1}(o_{h+1}|s_{h+1})| \\
&= \sum_{s_h, p_h} |\mathbb{P}_h^{\mathcal{G}}(s_h, p_h|c_h) - \mathbb{P}_h^{\mathcal{M},c}(s_h, p_h|\widehat{c}_h)|.
\end{aligned}
$$

Finally, since after marginalization, the total variation will not increase, we conclude that

$$\begin{aligned}
&\sum_{z_{h+1}} |\mathbb{P}_h^{\mathcal{G}}(z_{h+1} \,|\, c_h, \gamma_h) - \mathbb{P}_h^{\mathcal{M},z}(z_{h+1}|\widehat{c}_h, \gamma_h)| \\
&\leq \sum_{s_h,p_h,a_h,s_{h+1},o_{h+1}} |\mathbb{P}_h^{\mathcal{G}}(s_h, s_{h+1}, p_h, a_h, o_{h+1} \,|\, c_h, \gamma_h) - \mathbb{P}_h^{\mathcal{M}}(s_h, s_{h+1}, p_h, a_h, o_{h+1}|\widehat{c}_h, \gamma_h)|, \\
&\sum_{o_{h+1}} |\mathbb{P}_h^{\mathcal{G}}(o_{h+1} \,|\, c_h, \gamma_h) - \mathbb{P}_h^{\mathcal{M},o}(o_{h+1}|\widehat{c}_h, \gamma_h)| \\
&\leq \sum_{s_h,p_h,a_h,s_{h+1},o_{h+1}} |\mathbb{P}_h^{\mathcal{G}}(s_h, s_{h+1}, p_h, a_h, o_{h+1} \,|\, c_h, \gamma_h) - \mathbb{P}_h^{\mathcal{M}}(s_h, s_{h+1}, p_h, a_h, o_{h+1}|\widehat{c}_h, \gamma_h)|,
\end{aligned}$$

which proved the lemma. □

Therefore, in the following discussion, we only need to define $\widehat{c}_h$ and the corresponding belief $\{\mathbb{P}_h^{\mathcal{M},c}(s_h, p_h|\widehat{c}_h)\}_{h\in[H]}$. The definition of $\mathbb{P}_h^{\mathcal{M},z}(z_{h+1}|\widehat{c}_h, \gamma_h)$ and $\mathbb{E}^{\mathcal{M}}[\widehat{r}_{i,h+1}(o_{h+1})|\widehat{c}_h, \gamma_h]$ will follow from the consistency (B.3) and (B.4). Furthermore, it suffices to bound $\mathbb{E}_{a_{1:h-1},o_{2:h}\sim\pi'}\|\mathbb{P}_h^{\mathcal{G}}(\cdot,\cdot\,|\,c_h) - \mathbb{P}_h^{\mathcal{M},c}(\cdot,\cdot|\widehat{c}_h)\|_1$ since for the following discussion, we have been assuming knowledge of $\mathcal{G}$ and can just use true r for $\widehat{r}$. Now we will show when the information structure satisfies our Assumption 3 how we can construct approximate common information with history truncation that satisfies Definition 5.

One-step delayed information-sharing. For this, the information structure has $c_h = \{a_{1:h-1}, o_{2:h-1}\}$, $p_{i,h} = \{o_{i,h}\}$, $z_{h+1} = \{o_h, a_h\}$. Fix $L > 0$, we define the approximate common information as $\widehat{c}_h = \{a_{h-L:h-1}, o_{h-L+1:h-1}\}$. Furthermore, define the common information conditioned belief as $\mathbb{P}_h^{\mathcal{M},c}(s_h, p_h|\widehat{c}_h) = \boldsymbol{b}'_h(a_{h-L:h-1}, o_{h-L:h-1})(s_h)\mathbb{O}_h(o_h|s_h)$. Now we are ready to verify that it satisfies Definition 5.

- Obviously, it satisfies condition (5.1).
- Note that for any $c_h \in \mathcal{C}_h$:

$$\begin{aligned}
&\|\mathbb{P}_h^{\mathcal{G}}(\cdot,\cdot\,|\,c_h) - \mathbb{P}_h^{\mathcal{M},c}(\cdot,\cdot|\widehat{c}_h)\|_1 \\
&= \sum_{s_h,o_h} |\boldsymbol{b}_h(a_{1:h-1}, o_{2:h-1})(s_h)\mathbb{O}_h(o_h|s_h) - \boldsymbol{b}'_h(a_{h-L:h-1}, o_{h-L+1:h-1})(s_h)\mathbb{O}_h(o_h|s_h)| \\
&= \|\boldsymbol{b}_h(a_{1:h-1}, o_{2:h-1}) - \boldsymbol{b}'_h(a_{h-L:h-1}, o_{h-L+1:h-1})\|_1.
\end{aligned}$$

  Therefore, by setting $L \geq C\gamma^{-4}\log(\frac{S}{\epsilon})$, according to Equation (E.7) in Theorem E.4, we conclude that for any $\pi' \in \Pi^{\mathrm{gen}}, h \in [H]$:

$$\begin{aligned}
&\mathbb{E}_{a_{1:h-1},o_{2:h}\sim\pi'}\|\mathbb{P}_h^{\mathcal{G}}(\cdot,\cdot\,|\,c_h) - \mathbb{P}_h^{\mathcal{M},c}(\cdot,\cdot|\widehat{c}_h)\|_1 \\
&\leq \mathbb{E}_{a_{1:h-1},o_{2:h}\sim\pi'}\|\boldsymbol{b}_h(a_{1:h-1}, o_{2:h-1}) - \boldsymbol{b}'_h(a_{h-L:h-1}, o_{h-L+1:h-1})\|_1 \leq \epsilon.
\end{aligned}$$

  Therefore, conditions (5.2), (5.3) are satisfied with $\epsilon_r = \epsilon_z = \epsilon$.

Finally, to guarantee $\widehat{\pi}^\star$ is an ϵ-NE/CE/CCE, according to our Theorem 2, one needs $L \geq C\gamma^{-4}\log(\frac{SH}{\epsilon})$. Formally, we have the following theorem:

Theorem 11. Let $\epsilon, \gamma > 0$. Algorithm 1 given a γ-observable POSG of one-step delayed information sharing has time complexity $H(AO)^{C\gamma^{-4}\log\frac{SH}{\epsilon}}\,\mathrm{poly}(S, A, O, H, \frac{1}{\epsilon})$ for some universal constant $C > 0$.

Proof. It is obvious that $\widehat{C}_h = (AO)^L$ and $P_h = AO$, the polynomial dependence on S, H, A, and O comes from computing $\mathbb{P}_h^{\mathcal{M},c}(s_h, p_h|\widehat{c}_h)$ and equilibrium subroutines. □

State controlled by one controller with asymmetric delay sharing. The information structure is given as $c_h = \{o_{1,2:h}, o_{2,2:h-d}, a_{1,1:h-1}\}$, $p_{1,h} = \emptyset$, $p_{2,h} = \{o_{2,h-d+1:h}\}$. It is a little tricky to verify Assumption 3 and $\mathbb{P}_h^{\mathcal{G}}(s_h, p_h\,|\,c_h)$ can be computed as follows. Denote $\tau_{h-d} = \{a_{1:h-d-1}, o_{2:h-d}\}$, $f_a = \{a_{1,h-d:h-1}\}$, $f_o = \{o_{1,h-d+1:h}\}$. Now $\mathbb{P}_h^{\mathcal{G}}(s_h, p_h\,|\,c_h) =$

$\sum_{s_{h-d}} \mathbb{P}^{\mathcal{G}}(s_h, p_h \,|\, s_{h-d}, f_a, f_o)\mathbb{P}^{\mathcal{G}}(s_{h-d} \,|\, \tau_{h-d}, f_a, f_o)$. It is easy to see that $\mathbb{P}^{\mathcal{G}}_h(s_h, p_h \,|\, s_{h-d}, f_a, f_o)$ does not depend on the policy. For $\mathbb{P}^{\mathcal{G}}(s_{h-d} \,|\, \tau_{h-d}, f_a, f_o)$, the following holds

$$\mathbb{P}^{\mathcal{G}}(s_{h-d} \,|\, \tau_{h-d}, f_a, f_o) = \frac{\mathbb{P}^{\mathcal{G}}(s_{h-d}, f_a, f_o \,|\, \tau_{h-d})}{\sum_{s'_{h-d}} \mathbb{P}^{\mathcal{G}}(s'_{h-d}, f_a, f_o \,|\, \tau_{h-d})}.$$

Now note that

$$\mathbb{P}^{\mathcal{G}}(s_{h-d}, f_a, f_o \,|\, \tau_{h-d}) = \boldsymbol{b}_{h-d}(a_{1:h-d-1}, o_{2:h-d})(s_{h-d})\mathbb{P}^{\mathcal{G}}(a_{1,h-d} \,|\, \tau_{h-d})\mathbb{P}^{\mathcal{G}}(o_{1,h-d+1} \,|\, s_{h-d}, a_{1,h-d}) \cdots \mathbb{P}^{\mathcal{G}}(o_{1,h} \,|\, s_{h-d}, a_{1,h-d:h-1}).$$

Now let us use the notation $P(f_o \,|\, s_{s-d}, f_a) := \prod_{t=1}^{d} \mathbb{P}^{\mathcal{G}}(o_{1,h-d+t} \,|\, s_{h-d}, a_{1,h-d:h-d+t-1})$. Then it holds that $\sum_{f_o} P(f_o \,|\, s_{h-d}, f_a) = 1$, which suggests that the notation $P(f_o \,|\, s_{s-d}, f_a)$ can be understood as a conditional probability. With such notation, $\mathbb{P}^{\mathcal{G}}(s_{h-d} \,|\, \tau_{h-d}, f_a, f_o) = \frac{\boldsymbol{b}_{h-d}(a_{1:h-d-1}, o_{2:h-d})(s_{h-d})P(f_o \,|\, s_{h-d}, f_a)}{\sum_{s'_{h-d}} \boldsymbol{b}_{h-d}(a_{1:h-d-1}, o_{2:h-d})(s'_{h-d})P(f_o \,|\, s'_{h-d}, f_a)} = F^{P(\cdot\,|\,\cdot, f_a)}(\boldsymbol{b}_{h-d}(a_{1:h-d-1}, o_{2:h-d}); f_o)(s_{h-d})$. Finally, we compute:

$$\mathbb{P}^{\mathcal{G}}_h(s_h, p_h \,|\, c_h) = \sum_{s_{h-d}} \mathbb{P}^{\mathcal{G}}(s_h, p_h \,|\, s_{h-d}, f_a, f_o) F^{P(\cdot\,|\,\cdot, f_a)}(\boldsymbol{b}_{h-d}(a_{1:h-d-1}, o_{2:h-d}); f_o)(s_{h-d}).$$

Now for some fixed $L > 0$, we construct the approximate common information $\widehat{c}_h := \{o_{1,h-d-L+1:h}, o_{2,h-d-L+1:h-d}, a_{1,h-d-L:h-1}\}$ and correspondingly.

$$\mathbb{P}^{\mathcal{M},c}_h(s_h, p_h \,|\, \widehat{c}_h) = \sum_{s_{h-d}} \mathbb{P}^{\mathcal{G}}(s_h, p_h \,|\, s_{h-d}, f_a, f_o) F^{P(\cdot\,|\,\cdot, f_a)}(\boldsymbol{b}'_{h-d}(a_{h-d-L:h-d-1}, o_{h-d-L+1:h-d}); f_o)(s_{h-d}). \tag{E.14}$$

To verify Definition 5:

- Obviously, it satisfies the condition (5.1).
- For any $c_h \in \mathcal{C}_h$, it holds that

$$\begin{aligned}
&\|\mathbb{P}^{\mathcal{G}}_h(\cdot, \cdot \,|\, c_h) - \mathbb{P}^{\mathcal{M},c}_h(\cdot, \cdot \,|\, \widehat{c}_h)\|_1 \\
&\le \|F^{P(\cdot\,|\,\cdot, f_a)}(\boldsymbol{b}_{h-d}(a_{1:h-d-1}, o_{2:h-d}); f_o) - F^{P(\cdot\,|\,\cdot, f_a)}(\boldsymbol{b}'_{h-d}(a_{h-d-L:h-d-1}, o_{h-d-L+1:h-d}); f_o)\|_1.
\end{aligned}$$

  Finally, for any policy $\pi' \in \Pi^{\mathrm{gen}}$ taking expectations over τ_{h-d}, f_a, f_o, we conclude that as long as $L \ge C\gamma^{-4} \log \frac{S}{\epsilon}$, we conclude that

$$\mathbb{E}^{\mathcal{G}}_{a_{1:h-1}, o_{2:h} \sim \pi'} \|\mathbb{P}^{\mathcal{G}}_h(\cdot, \cdot \,|\, c_h) - \mathbb{P}^{\mathcal{M},c}_h(\cdot, \cdot \,|\, \widehat{c}_h)\|_1 \le \epsilon.$$

Finally, to guarantee $\widehat{\pi}^\star$ is ϵ-NE/CE/CCE, according to Theorem 2, one needs $L \ge C\gamma^{-4} \log(\frac{SH}{\epsilon})$. Formally, we have the following theorem:

Theorem 12. Let $\epsilon, \gamma > 0$. Algorithm 1 given a γ-observable POSG of state controlled by one controller with asymmetric delay sharing has time complexity $H(AO)^{C(\gamma^{-4} \log \frac{SH}{\epsilon} + d)} \mathrm{poly}(S, A, O, H, \frac{1}{\epsilon})$ for some universal constant $C > 0$.

Proof. It follows from the fact that $\widehat{C}_h \le (AO)^{L+d}$ and $P_h \le O_2^d$. The polynomial dependence on S, H, A, and O comes from computing $\mathbb{P}^{\mathcal{M},c}_h(s_h, p_h | \widehat{c}_h)$ and the equilibrium subroutines. □

Information sharing with one-directional-one-step delay. For this case, we have $c_h = \{a_{1:h-1}, o_{2:h-1}, o_{1,h}\}$, $p_{1,h} = \emptyset$, $p_{2,h} = \{o_{2,h}\}$, and $z_{h+1} = \{o_{1,h+1}, o_{2,h}, a_h\}$. Fix $L > 0$, we construct the approximate common information as $\widehat{c}_h = \{a_{h-L:h-1}, o_{h-L+1:h-1}, o_{1,h}\}$. Furthermore, define the belief as $\mathbb{P}^{\mathcal{M},c}_h(s_h, p_h | \widehat{c}_h) = \boldsymbol{b}'_h(a_{h-L:h-1}, o_{h-L+1:h-1}, o_{1,h})(s_h)\mathbb{P}_h(o_{2,h} | s_h, o_{1,h})$, where $\mathbb{P}_h(o_{2,h} | s_h, o_{1,h}) = \frac{\mathbb{O}_h(o_{1,h}, o_{2,h} | s_h)}{\sum_{o'_{2,h}} \mathbb{O}_h(o_{1,h}, o'_{2,h} | s_h)}$. Now we are ready to verify that Definition 5 is satisfied.

- Obviously, the condition 5.1 is satisfied.

- Note that for any $c_h \in \mathcal{C}_h$:

$$\begin{aligned}
&\|\mathbb{P}_h^{\mathcal{G}}(\cdot,\cdot\,|\,c_h) - \mathbb{P}_h^{\mathcal{M},c}(\cdot,\cdot\,|\widehat{c}_h)\|_1 \\
&= \sum_{s_h,o_{2,h}} |\boldsymbol{b}_h(a_{1:h-1},o_{2:h-1},o_{1,h})(s_h)\mathbb{P}_h(o_{2,h}|s_h,o_{1,h}) \\
&- \boldsymbol{b}'_h(a_{h-L:h-1},o_{h-L+1:h-1},o_{1,h})(s_h)\mathbb{P}_h(o_{2,h}|s_h,o_{1,h})| \\
&= \|\boldsymbol{b}_h(a_{1:h-1},o_{2:h-1},o_{1,h}) - \boldsymbol{b}'_h(a_{h-L:h-1},o_{h-L+1:h-1},o_{1,h})\|_1
\end{aligned}$$

  Therefore, by setting $L \geq C\gamma^{-4}\log(\frac{S}{\epsilon})$, according to (E.9) in Theorem E.4, we conclude that for any $\pi' \in \Pi^{\mathrm{gen}}$:

$$\begin{aligned}
&\mathbb{E}_{a_{1:h-1},o_{2:h}\sim\pi'}\|\mathbb{P}_h^{\mathcal{G}}(\cdot,\cdot\,|\,c_h) - \mathbb{P}_h^{\mathcal{M},c}(\cdot,\cdot\,|\widehat{c}_h)\|_1 \\
&\leq \mathbb{E}_{a_{1:h-1},o_{2:h}\sim\pi'}\|\boldsymbol{b}_h(a_{1:h-1},o_{2:h-1},o_{1,h}) - \boldsymbol{b}'_h(a_{h-L:h-1},o_{h-L+1:h-1},o_{1,h})\|_1 \leq \epsilon.
\end{aligned}$$

  Therefore, conditions (5.2), (5.3) are satisfied with $\epsilon_r = \epsilon_z = \epsilon$.

Finally, to guarantee $\widehat{\pi}^\star$ is an ϵ-NE/CE/CCE, according to Theorem 2, one needs $L \geq C\gamma^{-4}\log(\frac{SH}{\epsilon})$. Formally, we have the following theorem:

Theorem 13. Let $\epsilon, \gamma > 0$. Algorithm 1 given a γ-observable POSG of information sharing with one-directional-one-step delay has time complexity $H(AO)^{C\gamma^{-4}\log\frac{SH}{\epsilon}}\mathrm{poly}(S,A,O,H,\frac{1}{\epsilon})$ for some universal constant $C > 0$.

Proof. It is obvious that $\widehat{C}_h = (AO)^L$ and $P_h = O_2$. The polynomial dependence on S, H, A, and O comes from computing $\mathbb{P}_h^{\mathcal{M},c}(s_h, p_h|\widehat{c}_h)$ and the equilibrium subroutines. □

Uncontrolled state process with delayed sharing. As long as the state transition does not depend on the actions, Assumption 3 is satisfied. For convenience we consider the most general d-step delayed sharing information structure, where $d \geq 0$ and not necessarily $d = 1$ like in the one-step delayed information sharing structure. The information structure satisfies $c_h = \{o_{2:h-d}\}$, $p_{i,h} = \{o_{i,h-d+1:h}\}$, and $z_{h+1} = \{o_{h-d+1}\}$. Fix a $L \geq 0$, the approximate common information is $\widehat{c}_h = \{o_{h-d-L+1:h-d}\}$, the corresponding belief is $\mathbb{P}_h^{\mathcal{M},c}(s_h, p_h|\widehat{c}_h) = \sum_{s_{h-d}} \boldsymbol{b}'_{h-d}(o_{h-d-L+1:h-d})(s_{h-d})\mathbb{P}_h^{\mathcal{G}}(s_h, o_{h-d+1:h}|s_{h-d})$. Now we are ready to verify Definition 5.

- Obviously, the condition (5.1) is satisfied.
- Note that for any c_h:

$$\begin{aligned}
&\|\mathbb{P}_h^{\mathcal{G}}(\cdot,\cdot\,|\,c_h) - \mathbb{P}_h^{\mathcal{M},c}(\cdot,\cdot\,|\widehat{c}_h)\|_1 \\
&= \sum_{s_h,o_{h-d+1:h}} |\sum_{s_{h-d}} \boldsymbol{b}_{h-d}(o_{2:h-d})(s_{h-d})\mathbb{P}_h^{\mathcal{G}}(s_h, o_{h-d+1:h}|s_{h-d}) - \sum_{s_{h-d}} \boldsymbol{b}'_{h-d}(o_{h-d-L+1:h-d})(s_{h-d})\mathbb{P}_h^{\mathcal{G}}(s_h, o_{h-d+1:h}|s_{h-d})| \\
&= \sum_{s_h,o_{h-d+1:h}} |\sum_{s_{h-d}} (\boldsymbol{b}_{h-d}(o_{2:h-d})(s_{h-d}) - \boldsymbol{b}'_{h-d}(o_{h-d-L+1:h-d})(s_{h-d}))\mathbb{P}_h^{\mathcal{G}}(s_h, o_{h-d+1:h}|s_{h-d})| \\
&\leq \|\boldsymbol{b}_{h-d}(o_{2:h-d}) - \boldsymbol{b}'_{h-d}(o_{h-d-L+1:h-d})\|_1,
\end{aligned}$$

  where for the last step, we use Lemma 9. Therefore, by setting $L \geq C\gamma^{-4}\log(\frac{S}{\epsilon})$, according to (E.8) in Theorem E.4, we conclude that for any $\pi' \in \Pi^{\mathrm{gen}}$:

$$\begin{aligned}
&\mathbb{E}_{a_{1:h-1},o_{2:h}\sim\pi'}\|\mathbb{P}_h^{\mathcal{G}}(\cdot,\cdot\,|\,c_h) - \mathbb{P}_h^{\mathcal{M},c}(\cdot,\cdot\,|\widehat{c}_h)\|_1 \\
&\leq \mathbb{E}_{a_{1:h-1},o_{2:h}\sim\pi'}\|\boldsymbol{b}_{h-d}(o_{2:h-d}) - \boldsymbol{b}'_{h-d}(o_{h-d-L+1:h-d})\|_1 \leq \epsilon.
\end{aligned}$$

  This verifies the conditions (5.2), (5.3) with $\epsilon_r = \epsilon_z = \epsilon$.

Finally, to guarantee $\widehat{\pi}^\star$ is ϵ-NE/CE/CCE, according to our Theorem 2, one needs $L \geq C\gamma^{-4}\log(\frac{SH}{\epsilon})$. Formally, we have the following theorem:

Theorem 14. Let $\epsilon, \gamma > 0$. Algorithm 1 given a γ-observable POSG of uncontrolled state process has time complexity $H(O)^{C\gamma^{-4}\log\frac{SH}{\epsilon}}\mathrm{poly}(S,A,O^d,H,\frac{1}{\epsilon})$ for some universal constant $C > 0$.

Proof. It is obvious that $\widehat{C}_h = O^L$ and $P_h = O^d$, The polynomial dependence on S, A, H, and O^d comes from computing $\mathbb{P}_h^{\mathcal{M},c}(s_h, p_h|\widehat{c}_h)$ and the equilibirum subroutines. □

Symmetric information game. For symmetric information game, it has the following information structure. $c_h = \{a_{1:h-1}, o_{2:h}\}$, $p_{i,h} = \emptyset$, and $z_{h+1} = \{a_h, o_{h+1}\}$. Fix $L \geq 0$, we construct the approximate common information as $\widehat{c}_h = \{a_{h-L:h-1}, o_{h-L+1:h}\}$. Furthermore, we define the belief $\mathbb{P}_h^{\mathcal{M},c}(s_h, p_h|\widehat{c}_h) = \boldsymbol{b}'_h(a_{h-L:h-1}, o_{h-L+1:h})(s_h)$. Now we are ready to verify Definition 5.

- Obviously, it satisfies the condition (5.1).
- Note that for any $c_h \in \mathcal{C}_h$:

$$\|\mathbb{P}_h^{\mathcal{G}}(\cdot,\cdot|c_h) - \mathbb{P}_h^{\mathcal{M},c}(\cdot,\cdot|\widehat{c}_h)\|_1 = \|\boldsymbol{b}_h(a_{1:h-1}, o_{2:h}) - \boldsymbol{b}'_h(a_{h-L,h-1}, o_{h-L+1:h})\|_1.$$

  Therefore, by setting $L \geq C\gamma^{-4}\log(\frac{S}{\epsilon})$, according to (E.8) in Theorem E.4, we conclude that for any $\pi' \in \Pi^{\mathrm{gen}}$:

$$\mathbb{E}_{a_{1:h-1}, o_{2:h} \sim \pi'} \|\mathbb{P}_h^{\mathcal{G}}(\cdot,\cdot|c_h) - \mathbb{P}_h^{\mathcal{M},c}(\cdot,\cdot|\widehat{c}_h)\|_1 = \|\boldsymbol{b}_h(a_{1:h-1}, o_{2:h}) - \boldsymbol{b}'_h(a_{h-L,h-1}, o_{h-L+1:h})\|_1 \leq \epsilon.$$

  Therefore, the conditions (5.2) and (5.3) are satisfied with $\epsilon_r = \epsilon_z = \epsilon$.

Finally, to guarantee $\widehat{\pi}^\star$ is ϵ-NE/CE/CCE, according to Theorem 2, one needs $L \geq C\gamma^{-4}\log(\frac{SH}{\epsilon})$. Formally, we have the following theorem:

Theorem 15. Let $\epsilon, \gamma > 0$. Algorithm 1 given a γ-observable POSG of symmetric information has time complexity $H(AO)^{C\gamma^{-4}\log\frac{SH}{\epsilon}}\,\mathrm{poly}(S, A, H, O, \frac{1}{\epsilon})$ for some universal constant $C > 0$.

Proof. It is obvious that $\widehat{C}_h = (AO)^L$ and $P_h = A$, the polynomial dependence on S, H, A, and O comes from computing $\mathbb{P}_h^{\mathcal{M},c}(s_h, p_h|\widehat{c}_h)$ and equilibrium subroutines. □

Lemma 9. For any given sequence $\{x_i\}_{i=1}^m$ and $\{\{y_{i,j}\}_{i=1}^m\}_{j=1}^n$ such that $\sum_{j=1}^n |y_{i,j}| = 1$, $\forall i \in [m], j \in [n]$. The following holds

$$\sum_{j=1}^n |\sum_{i=1}^m x_i y_{i,j}| \leq \sum_{i=1}^n |x_i|.$$

Proof. Let $\boldsymbol{x} = (x_1, \cdots, x_m)^\top$, $\boldsymbol{y}_j = (y_{1,j}, \cdots, y_{m,j})^\top$, and $\boldsymbol{Y} = (\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n)$. Therefore, we have

$$\sum_{j=1}^n |\sum_{i=1}^m x_i y_{i,j}| = \sum_{j=1}^n |\boldsymbol{x}^\top \boldsymbol{y}_j| = \|\boldsymbol{Y}^\top \boldsymbol{x}\|_1 \leq \|\boldsymbol{Y}^\top\|_1 \|\boldsymbol{x}\|_1.$$

Note that $\|\boldsymbol{Y}^\top\|_1 = \|\boldsymbol{Y}\|_\infty = \max_i \sum_{j=1}^n |y_{i,j}| = 1$. Therefore, we conclude by

$$\sum_{j=1}^n |\sum_{i=1}^m x_i y_{i,j}| \leq \sum_{i=1}^n |x_i|.$$

□

E.5. Learning with approximate common information

Note for our previous planning algorithm, we have been assuming that we know the true model (transition dynamics and rewards) of the POSG $\mathcal{G}$, which avoids the issue of strategic explorations. For learning NE/CE/CCE in $\mathcal{G}$, one could treat $\mathcal{G}$ as a (fully-observable) Markov game on the state space of c_h. However, this formulation could be neither computationally or sample efficient because of the typical large space of common information. Therefore, we have to learn NE/CE/CCE in an approximation $\mathcal{M}$ with the state space of $\widehat{c}_h$ in Definition 5. However, the key problem is that we can only sample according to the model of $\mathcal{G}$ instead of $\mathcal{M}$. To circumvent this issue, similar to the idea of (Golowich et al., 2022a), the solution is

to construct $\widetilde{\mathcal{M}}(\pi^{1:H})$ for a sequence of H policies $\pi^{1:H} = (\pi^1, \cdots, \pi^H)$, where $\pi_h \in \Pi^{\text{gen}}$ for any $h \in [H]$ such that the transition and rewards of $\widetilde{\mathcal{M}}(\pi^{1:H})$ at each step h is defined by executing the policy π^h. Formally, Proposition 1 verifies that $\widetilde{\mathcal{M}}(\pi^{1:H})$ can be simulated by executing policies π^h at each step h in $\mathcal{G}$. Therefore, different from a generic $\mathcal{M}$ in Definition 5, to which we do not have algorithmic access, such a delicately designed transition dynamic and reward function allow us to actually simulate $\widetilde{\mathcal{M}}(\pi^{1:H})$ by executing policies $\pi^{1:H}$ in $\mathcal{G}$.

The next question is how to explore the *state space* $\{\widehat{\mathcal{C}}\}_{h\in[H+1]}$. It turns out that when such a state $\widehat{c}_h$ comes from a sequence of observations and actions, a uniform policy can be used to explore the state space (Efroni et al., 2022; Uehara et al., 2022). Formally, define the under-explored set of $\widehat{c}_h$ and $\widehat{c}_h \cup p_h$ under some policy π as follows.

Definition 14. For each $h \in [H]$, $\zeta > 0$, and a policy π, define the set $\mathcal{C}^{\text{low}}_{h,\zeta}(\pi) \subseteq \widehat{\mathcal{C}}_h$ as

$$\mathcal{C}^{\text{low}}_{h,\zeta}(\pi) := \{\widehat{c}_h \in \widehat{\mathcal{C}}_h : d^{\pi,\mathcal{G}}_{\mathcal{C},h}(\widehat{c}_h) \le \zeta\},$$

the set $\mathcal{V}^{\text{low}}_{h,\zeta}(\pi) \subseteq \mathcal{V}_h := \widehat{\mathcal{C}}_h \times \mathcal{P}_h$ as

$$\mathcal{V}^{\text{low}}_{h,\zeta}(\pi) := \{v_h \in \mathcal{V}_h : d^{\pi,\mathcal{G}}_{\mathcal{V},h}(v_h) \le \zeta\},$$

and the set $\mathcal{X}^{\text{low}}_{h,\zeta}(\pi) \subseteq \mathcal{X}_h := \mathcal{A}^{\min\{h,\widehat{L}\}} \times \mathcal{O}^{\min\{h,\widehat{L}\}}$ as

$$\mathcal{X}^{\text{low}}_{h,\zeta}(\pi) := \{x_h \in \mathcal{X}_h : d^{\pi,\mathcal{G}}_{\mathcal{X},h}(x_h) \le \zeta\},$$

where $d^{\pi,\mathcal{G}}_{\mathcal{C},h}(\widehat{c}_h) := \mathbb{P}^{\pi,\mathcal{G}}_h(\widehat{c}_h)$, $d^{\pi,\mathcal{G}}_{\mathcal{S},h}(s_h) := \mathbb{P}^{\pi,\mathcal{G}}_h(s_h)$, $d^{\pi,\mathcal{G}}_{\mathcal{V},h}(v_h) := \mathbb{P}^{\pi,\mathcal{G}}_h(v_h)$, and $d^{\pi,\mathcal{G}}_{\mathcal{X},h}(x_h) := \mathbb{P}^{\pi,\mathcal{G}}_h(x_h)$.

Now we shall relate the under-explored set of $\widehat{c}_h$ with the under-explored set of $s_{h'}$ for some $h' \in [H]$. Firstly, define the under-explored states under some policy π as

$$\mathcal{U}^{\mathcal{G}}_{\phi,h}(\pi) := \{s \in \mathcal{S} : d^{\pi,\mathcal{G}}_{\mathcal{S},h}(s) < \phi\}.$$

Then the following lemma holds.

Lemma 10. Fix any $\zeta > 0, \phi > 0, h \in [H]$. Consider any policy π, π', such that π' takes uniformly random actions at each step from $\max\{h - \widehat{L}, 1\}$ to h, each chosen independently of all previous states, actions, and observations. Then

$$d^{\pi,\mathcal{G}}_{\mathcal{C},h}(\mathcal{C}^{\text{low}}_{h,\zeta}(\pi')) \le \frac{A^{2\widehat{L}} O^{\widehat{L}} \zeta}{\phi} + \mathbb{1}[h > \widehat{L}] \cdot d^{\pi,\mathcal{G}}_{\mathcal{S},h-\widehat{L}}(\mathcal{U}^{\mathcal{G}}_{\phi,h-\widehat{L}}(\pi')). \tag{E.15}$$

Proof. Note that we have for each $\widehat{c}_h \in \widehat{\mathcal{C}}_h$

$$d^{\pi,\mathcal{G}}_{\mathcal{C},h}(\widehat{c}_h) = \sum_{x_h : \widehat{f}_h(x_h) = \widehat{c}_h} d^{\pi,\mathcal{G}}_{\mathcal{X},h}(x_h).$$

Therefore, we have

$$\sum_{\widehat{c}_h \notin \mathcal{C}^{\text{low}}_{h,\zeta}(\pi')} d^{\pi,\mathcal{G}}_{\mathcal{C},h}(\widehat{c}_h) = \sum_{\widehat{c}_h \notin \mathcal{C}^{\text{low}}_{h,\zeta}(\pi')} \sum_{x_h : \widehat{f}_h(x_h) = \widehat{c}_h} d^{\pi,\mathcal{G}}_{\mathcal{X},h}(x_h) = \sum_{\widehat{f}_h(x_h) \notin \mathcal{C}^{\text{low}}_{h,\zeta}(\pi')} d^{\pi,\mathcal{G}}_{\mathcal{X},h}(x_h) \ge \sum_{x_h \notin \mathcal{X}^{\text{low}}_{h,\zeta}(\pi')} d^{\pi,\mathcal{G}}_{\mathcal{X},h}(x_h).$$

This leads to that

$$d^{\pi,\mathcal{G}}_{\mathcal{C},h}(\mathcal{C}^{\text{low}}_{h,\zeta}(\pi')) \le d^{\pi,\mathcal{G}}_{\mathcal{X},h}(\mathcal{X}^{\text{low}}_{h,\zeta}(\pi')) \le \frac{A^{2\widehat{L}} O^{\widehat{L}} \zeta}{\phi} + \mathbb{1}[h > \widehat{L}] \cdot d^{\pi,\mathcal{G}}_{\mathcal{S},h-\widehat{L}}(\mathcal{U}^{\mathcal{G}}_{\phi,h-\widehat{L}}(\pi')),$$

where in the second inequality, we use Lemma 10.4 of (Golowich et al., 2022a). □

The next step is to learn the transition and reward $\mathbb{P}_h^{\widetilde{\mathcal{M}}(\pi^{1:H}),z}(z_{h+1}|\widehat{c}_h,\gamma_h)$, $\mathbb{P}_h^{\widetilde{\mathcal{M}}(\pi^{1:H}),o}(o_{h+1}|\widehat{c}_h,\gamma_h)$ of the model $\widetilde{\mathcal{M}}(\pi^{1:H})$, which is equivalent to $\mathbb{P}_h^{\pi^h,\mathcal{G}}(z_{h+1}|\widehat{c}_h,\gamma_h)$ and $\mathbb{P}_h^{\pi^h,\mathcal{G}}(o_{h+1}|\widehat{c}_h,\gamma_h)$ through executing policies $\pi^{1:H}$ in $\mathcal{G}$. The challenge here is that although γ_h serves as the actions for the approximate game $\widetilde{\mathcal{M}}(\pi^{1:H})$, it is not possible to enumerate all possible actions, since γ_h is indeed continuous, and even if we only consider all the deterministic γ_h, the number of all possible mappings from the private information to the real actions in $\mathcal{G}$ is still of the order A^{P_h}. Therefore, learning $\mathbb{P}_h^{\widetilde{\mathcal{M}}(\pi^{1:H}),z}(z_{h+1}|\widehat{c}_h,\gamma_h)$ by enumerating all possible $\widehat{c}_h$ and γ_h is not statistically efficient. To circumvent this issue, note the fact that $\mathbb{P}_h^{\widetilde{\mathcal{M}}(\pi^{1:H}),z}(z_{h+1}|\widehat{c}_h,\gamma_h)=\mathbb{P}_h^{\pi^h,\mathcal{G}}(z_{h+1}|\widehat{c}_h,\gamma_h)=\sum_{\chi_{h+1}(p_h,a_h,o_{h+1})=z_{h+1}}\mathbb{P}_h^{\pi^h,\mathcal{G}}(p_h,a_h,o_{h+1}|\widehat{c}_h,\gamma_h)$, and the same for $\mathbb{P}_h^{\widetilde{\mathcal{M}}(\pi^{1:H}),o}$. Further, notice the decomposition for $\mathbb{P}_h^{\pi^h,\mathcal{G}}(p_h,a_h,o_{h+1}|\widehat{c}_h,\gamma_h)$:

$$\mathbb{P}_h^{\pi^h,\mathcal{G}}(p_h,a_h,o_{h+1}|\widehat{c}_h,\gamma_h)=\mathbb{P}_h^{\pi^h,\mathcal{G}}(p_h|\widehat{c}_h)\Pi_{i=1}^n\gamma_{i,h}(a_{i,h}|p_{i,h})\mathbb{P}_h^{\pi^h,\mathcal{G}}(o_{h+1}|\widehat{c}_h,p_h,a_h).$$

Therefore, it suffices to learn $\mathbb{P}_h^{\pi^h,\mathcal{G}}(p_h|\widehat{c}_h,\gamma_h)$ and $\mathbb{P}_h^{\pi^h,\mathcal{G}}(o_{h+1}|\widehat{c}_h,p_h,a_h)$. Formally, the following algorithm learns an approximation $\widehat{\mathcal{M}}(\pi^{1:H})$ of $\widetilde{\mathcal{M}}(\pi^{1:H})$. The algorithm for constructing the approximation enjoys the following guarantee. Before stating the guarantees, based on the evolution, we define $\{f_h\}_{h\in[H+1]}$ and $\{g_h\}_{h\in[H+1]}$ as mappings that maps the joint history to common information and private information.

Lemma 11. Fix $\delta_1,\zeta_1,\zeta_2,\theta_1,\theta_2>0$. Suppose for all $h\in[H]$, π^h satisfies the pre-conditions of Lemma 10, then as long as $N_0\geq\max\{\frac{C(\max_h P_h+\log\frac{4H\max_h\widehat{C}_h}{\delta_1})}{\zeta_1\theta_1^2},\frac{CA(O+\log\frac{4H\max_h(\widehat{C}_hP_h)A}{\delta_1})}{\zeta_2\theta_2^2}\}$ for some sufficiently large constant C, with probability at least $1-\delta_1$, the following event $\mathcal{E}_1$ holds:

- For all $h\in[H]$, $\widehat{c}_h\notin\mathcal{C}^{\text{low}}_{h,\zeta_1}(\pi^h)$, we have that

$$\|\mathbb{P}_h^{\widehat{\mathcal{M}}(\pi^{1:H})}(\cdot|\widehat{c}_h)-\mathbb{P}_h^{\pi^h,\mathcal{G}}(\cdot|\widehat{c}_h)\|_1\leq\theta_1. \tag{E.16}$$

- For all $h\in[H]$, $(\widehat{c}_h,p_h)\notin\mathcal{V}^{\text{low}}_{h,\zeta}(\pi^h)$, $a_h\in\mathcal{A}$, we have that

$$\|\mathbb{P}_h^{\widehat{\mathcal{M}}(\pi^{1:H})}(\cdot|\widehat{c}_h,p_h,a_h)-\mathbb{P}_h^{\pi^h,\mathcal{G}}(\cdot|\widehat{c}_h,p_h,a_h)\|_1\leq\theta_2. \tag{E.17}$$

Proof. We will prove the Equation (E.16) first. Note for any trajectory k of Algorithm 5, the distribution of p_h^k conditioned on $\widehat{c}_h^k$ is exactly $\mathbb{P}_h^{\pi^h,\mathcal{G}}(\cdot|\widehat{c}_h^k)$.

Now consider any $\widehat{c}_h\notin\mathcal{C}^{\text{low}}_{h,\zeta_1}(\pi^h)$. By Chernoff bound, with probability at least $1-\exp(-\frac{\zeta_1N_0}{8})$, there are at least $\frac{\zeta_1N_0}{2}$ trajectories $k\in[N_0]$, such that $\text{Compress}_h(f_h(a^k_{1:h-1},o^k_{2:h}))=\widehat{c}_h$. By the folklore theorem of learning a discrete probability distribution (Canonne, 2020), with probability $1-p'$, (E.16) holds as long as

$$\frac{\zeta_1N_0}{2}\geq\frac{C(P_h+\log\frac{1}{p'})}{\theta_1^2}, \tag{E.18}$$

for some constant $C>1$. By a union bound over all possible $h\in[H]$, and $\widehat{c}_h\in\widehat{\mathcal{C}}_h$, (E.16) holds with probability at least

$$1-H\max_h\widehat{C}_h\exp(-\frac{\zeta_1N_0}{8})-H\max_h\widehat{C}_hp'.$$

Now set $p'=\frac{\delta_1}{4H\max_h\widehat{C}_h}$ and it's easy to verify that (E.18) holds since $N_0\geq\frac{C(\max_h P_h+\log\frac{4H\max_h\widehat{C}_h}{\delta_1})}{\zeta_1\theta_1^2}$. Furthermore, as long as C is sufficiently large, we have that $H\max_h\widehat{C}_h\exp(-\frac{\zeta_1N_0}{8})\leq\frac{\delta_1}{4}$. Therefore, we proved that with probability at least $1-\frac{\delta_1}{2}$, equation (E.16) holds for all $h\in[H]$, and $\widehat{c}_h\notin\mathcal{C}^{\text{low}}_{h,\zeta_1}(\pi^h)$.

Similarly, consider any trajectory k, the distribution of o_{h+1} conditioned on $\widehat{c}_h^k,p_h,a_h$ is exactly $\mathbb{P}_h^{\pi^h,\mathcal{G}}(\cdot|\widehat{c}_h,p_h,a_h)$. Now consider any $(\widehat{c}_h,p_h)\notin\mathcal{C}^{\text{low}}_{h,\zeta_2}(\pi^h)$ and $a_h\in\mathcal{A}$. Now note due to the assumption for π^h, it holds that $\mathbb{P}_h^{\pi^h,\mathcal{G}}(\widehat{c}_h,p_h,a_h)=$

$\mathbb{P}_h^{\pi^h,\mathcal{G}}(\widehat{c}_h, p_h)\mathbb{P}_h^{\pi^h,\mathcal{G}}(a_h|\widehat{c}_h, p_h) \geq \frac{\zeta_2}{A}$. By Chernoff bound, with probability at least $1-\exp(-\frac{\zeta_1 N_0}{8A})$, there are at least $\frac{\zeta_2 N_0}{2A}$ trajectories $k \in [N_0]$, such that $\text{Compress}_h(f_h(a^k_{1:h-1}, o^k_{2:h})) = \widehat{c}_h, g_h(a^k_{1:h-1}, o^k_{2:h}) = p_h, a^k_h = a_h$. Again with probability at least $1-p'$, (E.17) holds as long as

$$\frac{\zeta_2 N_0}{2A} \geq \frac{C(O + \log\frac{1}{p'})}{\theta_2^2},$$

for some constant $C \geq 1$. By a union bound over all possible $h \in [H], \widehat{c}_h, p_h, a_h$, (E.17) holds with probability at least

$$1 - H\max_h(\widehat{C}_h P_h)A\exp(-\frac{\zeta_2 N_0}{8A}) - H\max_h(\widehat{C}_h P_h)Ap'.$$

Now we set $p' = \frac{\delta_1}{4H\max_h(\widehat{C}_h P_h)A}$. Then since $N_0 > \frac{CA(O+\log\frac{4H\max_h(\widehat{C}_h P_h)A}{\delta_1})}{\zeta_2\theta_2^2}$, it holds that $H\max_h(\widehat{C}_h P_h)A\exp(-\frac{\zeta_2 N_0}{8A}) \leq \frac{\delta_1}{4}$ and $H\max_h(\widehat{C}_h P_h)Ap' \leq \frac{\delta_1}{4}$ as long as the constant C is sufficiently large. Therefore, we conclude that with probability at least $1-\frac{\delta_1}{2}$, equation (E.17) holds for all $h \in [H], \widehat{c}_h \in \widehat{\mathcal{C}}_h, p_h \in \mathcal{P}_h, a_h \in \mathcal{A}$. Finally, by a union bound, we proved the lemma. □

With the previous lemma, the next step is to bound the two important quantity in Definition 5. In the following discussion, we will use $\widetilde{\mathcal{M}}$ for $\widetilde{\mathcal{M}}(\pi^{1:H})$, and $\widehat{\mathcal{M}}$ for $\widehat{\mathcal{M}}(\pi^{1:H})$.

Lemma 12. Under the event $\mathcal{E}_1$ in Lemma 11, for any $h \in [H]$, policy $\pi \in \Pi^{\text{gen}}$, reward function $\widehat{r}_{i,h} : \mathcal{O} \to [0,1]$ for any $i \in [n]$, $h \in [H+1]$, and prescription $\gamma_h \in \Gamma_h$, it holds that

$$\mathbb{E}^{\mathcal{G}}_{a_{1:h-1}, o_{2:h}\sim\pi} \sum_{z_{h+1}} |\mathbb{P}_h^{\widetilde{\mathcal{M}},z}(z_{h+1}\,|\widehat{c}_h, \gamma_h) - \mathbb{P}_h^{\widehat{\mathcal{M}},z}(z_{h+1}\,|\widehat{c}_h, \gamma_h)| \leq$$
$$O\theta_1 + 2AP_h\frac{\zeta_2}{\zeta_1} + AP_h\theta_2 + \frac{A^{2\widehat{L}}O^{\widehat{L}}\zeta_1}{\phi} + \mathbb{1}[h > \widehat{L}]\cdot 2\cdot d^{\pi,\mathcal{G}}_{\mathcal{S},h-\widehat{L}}(\mathcal{U}^{\mathcal{G}}_{\phi,h-\widehat{L}}(\pi^h)), \tag{E.19}$$
$$\mathbb{E}^{\mathcal{G}}_{a_{1:h-1}, o_{2:h}\sim\pi} |\mathbb{E}^{\widetilde{\mathcal{M}}}[\widehat{r}_{i,h+1}(o_{h+1})\,|\,\widehat{c}_h, \gamma_h] - \mathbb{E}^{\widehat{\mathcal{M}}}[\widehat{r}_{i,h+1}(o_{h+1})\,|\,\widehat{c}_h, \gamma_h]| \leq$$
$$O\theta_1 + 2AP_h\frac{\zeta_2}{\zeta_1} + AP_h\theta_2 + \frac{A^{2\widehat{L}}O^{\widehat{L}}\zeta_1}{\phi} + \mathbb{1}[h > \widehat{L}]\cdot 2\cdot d^{\pi,\mathcal{G}}_{\mathcal{S},h-\widehat{L}}(\mathcal{U}^{\mathcal{G}}_{\phi,h-\widehat{L}}(\pi^h)). \tag{E.20}$$

Proof. Note that

$$|\mathbb{E}^{\widetilde{\mathcal{M}}}[\widehat{r}_{i,h+1}(o_{h+1})\,|\,\widehat{c}_h, \gamma_h] - \mathbb{E}^{\widehat{\mathcal{M}}}[\widehat{r}_{i,h+1}(o_{h+1})\,|\,\widehat{c}_h, \gamma_h]| \leq \sum_{o_{h+1}} |\mathbb{P}_h^{\widetilde{\mathcal{M}},o}(o_{h+1}\,|\widehat{c}_h, \gamma_h) - \mathbb{P}_h^{\widehat{\mathcal{M}},o}(o_{h+1}|\widehat{c}_h, \gamma_h)|.$$

Therefore, to prove (E.20), it suffices to bound $\sum_{o_{h+1}} |\mathbb{P}_h^{\widetilde{\mathcal{M}},o}(o_{h+1}\,|\widehat{c}_h, \gamma_h) - \mathbb{P}_h^{\widehat{\mathcal{M}},o}(o_{h+1}|\widehat{c}_h, \gamma_h)|$. Under the event $\mathcal{E}_1$, consider

any $\widehat{c}_h \notin \mathcal{C}^{\text{low}}_{h,\zeta_1}(\pi^h)$ and $\gamma_h \in \Gamma_h$:

$$
\begin{aligned}
&\sum_{p_h,a_h,o_{h+1}} |\mathbb{P}_h^{\widetilde{\mathcal{M}}}(p_h,a_h,o_{h+1}|\widehat{c}_h,\gamma_h) - \mathbb{P}_h^{\widehat{\mathcal{M}}}(p_h,a_h,o_{h+1}|\widehat{c}_h,\gamma_h)| \\
&= \sum_{p_h,a_h,o_{h+1}} |\mathbb{P}_h^{\pi^h,\mathcal{G}}(p_h|\widehat{c}_h)\Pi_{i=1}^n \gamma_{i,h}(a_{i,h}\,|\,p_{i,h})\mathbb{P}_h^{\pi^h,\mathcal{G}}(o_{h+1}|\widehat{c}_h,p_h,a_h) - \mathbb{P}_h^{\widehat{\mathcal{M}}}(p_h|\widehat{c}_h)\Pi_{i=1}^n \gamma_{i,h}(a_{i,h}\,|\,p_{i,h})\mathbb{P}_h^{\widehat{\mathcal{M}},o}(o_{h+1}|\widehat{c}_h,p_h,a_h)| \\
&\le \sum_{p_h,a_h,o_{h+1}} \Pi_{i=1}^n \gamma_{i,h}(a_{i,h}\,|\,p_{i,h})|\mathbb{P}_h^{\pi^h,\mathcal{G}}(p_h|\widehat{c}_h) - \mathbb{P}_h^{\widehat{\mathcal{M}}}(p_h|\widehat{c}_h)| + \\
&\quad \Pi_{i=1}^n \gamma_{i,h}(a_{i,h}\,|\,p_{i,h})\mathbb{P}_h^{\pi^h,\mathcal{G}}(p_h|\widehat{c}_h)|\mathbb{P}_h^{\pi^h,\mathcal{G}}(o_{h+1}|\widehat{c}_h,p_h,a_h) - \mathbb{P}_h^{\widehat{\mathcal{M}},o}(o_{h+1}|\widehat{c}_h,p_h,a_h)| \\
&\le O\|\mathbb{P}_h^{\pi^h,\mathcal{G}}(\cdot|\widehat{c}_h) - \mathbb{P}_h^{\widehat{\mathcal{M}}}(\cdot|\widehat{c}_h)\|_1 + \sum_{p_h,a_h} \mathbb{P}_h^{\pi^h,\mathcal{G}}(p_h|\widehat{c}_h)\|\mathbb{P}_h^{\pi^h,\mathcal{G}}(\cdot|\widehat{c}_h,p_h,a_h) - \mathbb{P}_h^{\widehat{\mathcal{M}}}(\cdot|\widehat{c}_h,p_h,a_h)\|_1 \\
&\le O\theta_1 + \left(\sum_{p_h:\mathbb{P}_h^{\pi^h,\mathcal{G}}(p_h|\widehat{c}_h)\le\frac{\zeta_2}{\zeta_1}} + \sum_{p_h:\mathbb{P}_h^{\pi^h,\mathcal{G}}(p_h|\widehat{c}_h)>\frac{\zeta_2}{\zeta_1}}\right)\sum_{a_h} \mathbb{P}_h^{\pi^h,\mathcal{G}}(p_h|\widehat{c}_h)\|\mathbb{P}_h^{\pi^h,\mathcal{G}}(\cdot|\widehat{c}_h,p_h,a_h) - \mathbb{P}_h^{\widehat{\mathcal{M}}}(\cdot|\widehat{c}_h,p_h,a_h)\|_1 \\
&\le O\theta_1 + 2AP_h\frac{\zeta_2}{\zeta_1} + AP_h\theta_2,
\end{aligned}
$$

where the last inequality comes from the fact that if $\widehat{c}_h \notin \mathcal{C}^{\text{low}}_{h,\zeta_1}(\pi^h)$ and $\mathbb{P}_h^{\pi^h,\mathcal{G}}(p_h|\widehat{c}_h) > \frac{\zeta_2}{\zeta_1}$, then $(\widehat{c}_h,p_h) \notin \mathcal{V}^{\text{low}}_{h,\zeta_2}(\pi^h)$. Finally, for any policy π, by taking expectations over $\widehat{c}_h$, we conclude that

$$
\begin{aligned}
&\mathbb{E}^{\mathcal{G}}_{a_{1:h-1},o_{2:h}\sim\pi} \sum_{p_h,a_h,o_{h+1}} |\mathbb{P}_h^{\widetilde{\mathcal{M}}}(p_h,a_h,o_{h+1}|\widehat{c}_h,\gamma_h) - \mathbb{P}_h^{\widehat{\mathcal{M}}}(p_h,a_h,o_{h+1}|\widehat{c}_h,\gamma_h)| \\
&\le O\theta_1 + 2AP_h\frac{\zeta_2}{\zeta_1} + AP_h\theta_2 + 2\cdot d^{\pi,\mathcal{G}}_{\mathcal{C},h}(\mathcal{C}^{\text{low}}_{h,\zeta_1}(\pi^h)) \\
&\le O\theta_1 + 2AP_h\frac{\zeta_2}{\zeta_1} + AP_h\theta_2 + \frac{A^{2\widehat{L}}O^{\widehat{L}}\zeta_1}{\phi} + \mathbb{1}[h>\widehat{L}]\cdot 2\cdot d^{\pi,\mathcal{G}}_{\mathcal{S},h-\widehat{L}}(\mathcal{U}^{\mathcal{G}}_{\phi,h-\widehat{L}}(\pi^h)),
\end{aligned}
$$

where the last step comes from Lemma 10. By noticing that after marginalization the total variation will not increase:

$$
\begin{aligned}
&\sum_{z_{h+1}} |\mathbb{P}_h^{\widetilde{\mathcal{M}},z}(z_{h+1}\,|\widehat{c}_h,\gamma_h) - \mathbb{P}_h^{\widehat{\mathcal{M}},z}(z_{h+1}\,|\widehat{c}_h,\gamma_h)| \le \sum_{p_h,a_h,o_{h+1}} |\mathbb{P}_h^{\widetilde{\mathcal{M}}}(p_h,a_h,o_{h+1}|\widehat{c}_h,\gamma_h) - \mathbb{P}_h^{\widehat{\mathcal{M}}}(p_h,a_h,o_{h+1}|\widehat{c}_h,\gamma_h)| \\
&\sum_{o_{h+1}} |\mathbb{P}_h^{\widetilde{\mathcal{M}},o}(o_{h+1}\,|\widehat{c}_h,\gamma_h) - \mathbb{P}_h^{\widehat{\mathcal{M}},o}(o_{h+1}|\widehat{c}_h,\gamma_h)| \le \sum_{p_h,a_h,o_{h+1}} |\mathbb{P}_h^{\widetilde{\mathcal{M}}}(p_h,a_h,o_{h+1}|\widehat{c}_h,\gamma_h) - \mathbb{P}_h^{\widehat{\mathcal{M}}}(p_h,a_h,o_{h+1}|\widehat{c}_h,\gamma_h)|,
\end{aligned}
$$

we proved the lemma. □

Finally, we are ready to prove Theorem 8, building the relationship between $\mathcal{G}$ and $\widehat{\mathcal{M}}(\pi^{1:H})$ through $\widetilde{\mathcal{M}}(\pi^{1:H})$.

Proof. In the following proof, we will use $\widetilde{\mathcal{M}}$ for $\widetilde{\mathcal{M}}(\pi^{1:H})$ and $\widehat{\mathcal{M}}$ for $\widehat{\mathcal{M}}(\pi^{1:H})$. Note that for $\epsilon_r(\widehat{\mathcal{M}},\widehat{r})$, it holds that

$$
\begin{aligned}
\epsilon_r(\widehat{\mathcal{M}},\widehat{r}) &= \max_h \max_{\pi,\gamma_h} \mathbb{E}^{\mathcal{G}}_{a_{1:h-1},o_{2:h}\sim\pi} |\mathbb{E}^{\mathcal{G}}[r_{i,h+1}(o_{h+1})\mid c_h,\gamma_h] - \mathbb{E}^{\widehat{\mathcal{M}}}[\widehat{r}_{i,h+1}(o_{h+1})\mid\widehat{c}_h,\gamma_h]| \\
&\le \max_h \max_{\pi,\gamma_h} \mathbb{E}^{\mathcal{G}}_{a_{1:h-1},o_{2:h}\sim\pi} |\mathbb{E}^{\mathcal{G}}[r_{i,h+1}(o_{h+1})\mid c_h,\gamma_h] - \mathbb{E}^{\widetilde{\mathcal{M}}}[\widehat{r}_{i,h+1}(o_{h+1})\mid\widehat{c}_h,\gamma_h]| \\
&\quad + \max_h \max_{\pi,\gamma_h} \mathbb{E}^{\mathcal{G}}_{a_{1:h-1},o_{2:h}\sim\pi} |\mathbb{E}^{\widetilde{\mathcal{M}}}[\widehat{r}_{i,h+1}(o_{h+1})\mid c_h,\gamma_h] - \mathbb{E}^{\widehat{\mathcal{M}}}[\widehat{r}_{i,h+1}(o_{h+1})\mid\widehat{c}_h,\gamma_h]| \\
&\le \epsilon_r(\pi^{1:H},\widehat{r}) + \epsilon_{apx}(\pi^{1:H}),
\end{aligned}
$$

where the last step comes from Lemma 12. Similarly, for $\epsilon_z(\widehat{\mathcal{M}})$, it holds that

$$\begin{aligned}
\epsilon_z(\widehat{\mathcal{M}}) &= \max_h \max_{\pi,\gamma_h} \mathbb{E}^{\mathcal{G}}_{a_{1:h-1},o_{2:h}\sim\pi} \|\mathbb{P}^{\mathcal{G}}_h(\cdot\,|\,c_h,\gamma_h) - \mathbb{P}^{\widehat{\mathcal{M}},z}_h(\cdot\,|\,\widehat{c}_h,\gamma_h)\|_1 \\
&\le \max_h \max_{\pi,\gamma_h} \mathbb{E}^{\mathcal{G}}_{a_{1:h-1},o_{2:h}\sim\pi} \|\mathbb{P}^{\mathcal{G}}_h(\cdot\,|\,c_h,\gamma_h) - \mathbb{P}^{\widetilde{\mathcal{M}},z}_h(\cdot\,|\,\widehat{c}_h,\gamma_h)\|_1 \\
&\quad + \max_h \max_{\pi,\gamma_h} \mathbb{E}^{\mathcal{G}}_{a_{1:h-1},o_{2:h}\sim\pi} \|\mathbb{P}^{\widetilde{\mathcal{M}},z}_h(\cdot\,|\,c_h,\gamma_h) - \mathbb{P}^{\widehat{\mathcal{M}},z}_h(\cdot\,|\,\widehat{c}_h,\gamma_h)\|_1 \\
&\le \epsilon_z(\pi^{1:H}) + \epsilon_{apx}(\pi^{1:H})
\end{aligned}$$

where the last step again comes from Lemma 12. Therefore, with Lemma 4 and Theorem 2, we proved the Theorem. □

E.6. Learning with finite memory as approximate common information

Until now, we have not considered the relationship between $\widetilde{\mathcal{M}}(\pi^{1:H})$ and $\mathcal{G}$, which will necessarily depend on the choice of approximate common information $\widehat{c}_h$ and $\pi^{1:H}$. For planning, we have seen how to construct an approximate $\widehat{c}_h$ using finite memory. Similarly, here we will also show how to construct $\widehat{c}_h$ with finite memory so that $\widehat{\mathcal{M}}(\pi^{1:H})$ is a good approximation of $\mathcal{G}$. In the following discussions, we shall use another important policy-dependent approximate belief $\widetilde{\boldsymbol{b}}^{\pi}_h(a_{h-L:h-1}, o_{h-L+1:h}) := \boldsymbol{b}^{\mathrm{apx},\mathcal{G}}_h(a_{h-L:h-1}, o_{h-L+1:h}; d^{\pi,\mathcal{G}}_{\mathcal{S},h-L})$.We shall need the following important lemmas.

Lemma 13. There is a constant $C \ge 1$ so that the following holds. If Assumption 2 holds, then for any $\epsilon, \phi > 0, L \in \mathbb{N}$ so that $L \ge C\gamma^{-4}\log(\frac{1}{\epsilon\phi})$, it holds that for any policies $\pi \in \Pi$, $\pi' \in \Pi^{\mathrm{gen}}$,

$$\begin{aligned}
\mathbb{E}^{\mathcal{G}}_{a_{1:h-1},o_{2:h}\sim\pi'} \left\|\boldsymbol{b}_h(a_{1:h-1}, o_{2:h}) - \widetilde{\boldsymbol{b}}^{\pi}_h(a_{h-L:h-1}, o_{h-L+1:h})\right\|_1 &\le \epsilon + \mathbb{1}[h > L]\cdot 6\cdot d^{\pi',\mathcal{G}}_{\mathcal{S},h-L}\left(\mathcal{U}^{\mathcal{G}}_{\phi,h-L}(\pi)\right), \\
\mathbb{E}^{\mathcal{G}}_{a_{1:h-1},o_{2:h}\sim\pi'} \left\|\boldsymbol{b}_h(a_{1:h-1}, o_{2:h-1}) - \widetilde{\boldsymbol{b}}^{\pi}_h(a_{h-L:h-1}, o_{h-L+1:h-1})\right\|_1 &\le \epsilon + \mathbb{1}[h > L]\cdot 6\cdot d^{\pi',\mathcal{G}}_{\mathcal{S},h-L}\left(\mathcal{U}^{\mathcal{G}}_{\phi,h-L}(\pi)\right), \\
\mathbb{E}^{\mathcal{G}}_{a_{1:h-1},o_{2:h}\sim\pi'} \left\|\boldsymbol{b}_h(a_{1:h-1}, o_{2:h-1}, o_{1,h}) - \widetilde{\boldsymbol{b}}^{\pi}_h(a_{h-L:h-1}, o_{h-L+1:h-1}, o_{1,h})\right\|_1 &\le \epsilon + \mathbb{1}[h > L]\cdot 6\cdot d^{\pi',\mathcal{G}}_{\mathcal{S},h-L}\left(\mathcal{U}^{\mathcal{G}}_{\phi,h-L}(\pi)\right).
\end{aligned}$$

Furthermore, for any finite domain Y, conditional probability $q(y\,|\,s)$, and the posterior update operator $F^q : \Delta(\mathcal{S}) \to \Delta(\mathcal{S})$ as defined in Lemma 6, it holds that

$$\mathbb{E}^{\mathcal{G}}_{a_{1:h-1},o_{2:h}\sim\pi'} \mathbb{E}_{y\sim q\cdot\boldsymbol{b}_h(a_{1:h-1},o_{2:h})} \|F^q(\boldsymbol{b}_h(a_{1:h-1}, o_{2:h}); y) - F^q(\boldsymbol{b}'_h(a_{h-L:h-1}, o_{h-L+1:h}); y)\|_1 \le \epsilon.$$

Proof. It directly follows from Theorem E.4 and Lemma 12.2 in (Golowich et al., 2022a). □

Note the lemma shows that if we use the $d^{\pi,\mathcal{G}}_{\mathcal{S},h-\widehat{L}}$ instead of a $\mathrm{Unif}(\mathcal{S})$ as the prior, the approximate belief will suffer from an additional error term $d^{\pi',\mathcal{G}}_{\mathcal{S},h-L}\left(\mathcal{U}^{\mathcal{G}}_{\phi,h-L}(\pi)\right)$. The following lemma shows there already exists an efficient algorithm for finding π to minimize $d^{\pi',\mathcal{G}}_{\mathcal{S},h-L}\left(\mathcal{U}^{\mathcal{G}}_{\phi,h-L}(\pi)\right)$.

Lemma 14. Given $\alpha, \beta > 0$, $\widehat{L} \ge C\frac{\log(HSO/(\alpha\gamma))}{\gamma^4}$, and $\phi = \frac{\alpha\gamma^2}{C^3H^{10}S^5O^4}$ for some constant $C > 0$. There exists an algorithm `BaSeCAMP` with both computation and sample complexity bounded by $(OA)^{\widehat{L}}\log(\frac{1}{\beta})$ outputting $K = 2HS$ groups of policies $\{\pi^{1:H,j}\}_{j=1}^K$, where $\pi^{h,j}_{h'} = \mathrm{Unif}(\mathcal{A})$ for $h' \ge h - \widehat{L}, j \in [K]$ and rewards $\{(\widehat{r}^j_i)_{i=1}^n\}_{j=1}^K$. It holds that with probability at least $1-\beta$, there is at least one $j^\star \in [K]$ such that for any $h > \widehat{L}$, policy $\pi \in \Pi^{\mathrm{gen}}$

$$\begin{aligned}
d^{\pi,\mathcal{G}}_{\mathcal{S},h-\widehat{L}}(\mathcal{U}^{\mathcal{G}}_{\phi,h-\widehat{L}}(\pi^{h,j^\star})) &\le \frac{\alpha}{CH^2}, \\
\mathbb{E}^{\mathcal{G}}_{a_{1:h-1},o_{1:h}\sim\pi} |r_{i,h}(o_h) - \widehat{r}^{j^\star}_{i,h}(o_h)| &\le \frac{\alpha}{CH^2}.
\end{aligned}$$

Proof. It follows from Theorem 3.1 in (Golowich et al., 2022a). □

By combining two previous lemmas, we can show the following corollary:

Corollary 2. Given $\epsilon, \delta_2 > 0$, $L \geq C\frac{\log(HSO/(\epsilon\gamma))}{\gamma^4}$, and $\phi = \frac{\epsilon\gamma^2}{C^2H^8S^5O^4}$ for some constant $C > 0$. There exists an algorithm BaSeCAMP with both computation and sample complexity bounded by $N_1 = (OA)^L \log(\frac{1}{\delta_2})$ outputting $K = 2HS$ groups of policies $\{\pi^{1:H,j}\}_{j=1}^K$, where $\pi_{h'}^{h,j} = \mathrm{Unif}(\mathcal{A})$ for $h' \geq h - L, j \in [K]$. The following event $\mathcal{E}_2$ holds with probability at least $1 - \delta_2$: there is at least one $j^\star \in [K]$ such that for any $h > L$, policy $\pi' \in \Pi^{\mathrm{gen}}$

$$\mathbb{E}^{\mathcal{G}}_{a_{1:h-1}, o_{2:h} \sim \pi'} \left\| \boldsymbol{b}_h(a_{1:h-1}, o_{2:h}) - \widetilde{\boldsymbol{b}}_h^{\pi^{h,j^\star}}(a_{h-L:h-1}, o_{h-L+1:h}) \right\|_1 \leq \epsilon + \mathbb{1}[h > L] \cdot 6 \cdot d^{\pi', \mathcal{G}}_{\mathcal{S}, h-L}\left(\mathcal{U}^{\mathcal{G}}_{\phi, h-L}\left(\pi^{h,j^\star}\right)\right),$$

$$\mathbb{E}^{\mathcal{G}}_{a_{1:h-1}, o_{2:h} \sim \pi'} \left\| \boldsymbol{b}_h(a_{1:h-1}, o_{2:h-1}) - \widetilde{\boldsymbol{b}}_h^{\pi^{h,j^\star}}(a_{h-L:h-1}, o_{h-L+1:h-1}) \right\|_1 \leq \epsilon + \mathbb{1}[h > L] \cdot 6 \cdot d^{\pi', \mathcal{G}}_{\mathcal{S}, h-L}\left(\mathcal{U}^{\mathcal{G}}_{\phi, h-L}\left(\pi^{h,j^\star}\right)\right),$$

$$\mathbb{E}^{\mathcal{G}}_{a_{1:h-1}, o_{2:h} \sim \pi'} \left\| \boldsymbol{b}_h(a_{1:h-1}, o_{2:h-1}, o_{i,h}) - \widetilde{\boldsymbol{b}}_h^{\pi^{h,j^\star}}(a_{h-L:h-1}, o_{h-L+1:h-1}, o_{i,h}) \right\|_1 \leq \epsilon + \mathbb{1}[h > L] \cdot 6 \cdot d^{\pi', \mathcal{G}}_{\mathcal{S}, h-L}\left(\mathcal{U}^{\mathcal{G}}_{\phi, h-L}\left(\pi^{h,j^\star}\right)\right),$$

$$d^{\pi', \mathcal{G}}_{\mathcal{S}, h-L}(\mathcal{U}^{\mathcal{G}}_{\phi, h-L}(\pi^{h,j^\star})) \leq \epsilon,$$

$$\mathbb{E}^{\mathcal{G}}_{a_{1:h-1}, o_{2:h} \sim \pi} |r_{i,h}(o_h) - \widehat{r}^{j^\star}_{i,h}(o_h)| \leq \epsilon.$$

Proof. Let $\alpha = \frac{CH^2\epsilon}{2}$, $\delta_2 = \beta$, and $L \geq \max\{C\frac{\log(\frac{1}{\epsilon\phi})}{\gamma^4}, C\frac{\log(HSO/(\alpha\gamma))}{\gamma^4}\}$. Combing Lemma 13 and 14 leads to the conclusion. □

In the discussion thereafter, we will use $\widetilde{\mathcal{M}}$ for $\widetilde{\mathcal{M}}(\pi^{1:H,j^\star})$ and $\widehat{\mathcal{G}}$ for $\widehat{\mathcal{M}}(\pi^{1:H,j^\star})$, and $\widehat{r}_{i,h}$ for $\widehat{r}^{j^\star}_{i,h}$ interchangeably. There is still one issue unsolved, which is that BaSeCAMP does not tell us which $j \in [K]$ is the $j^\star$ we want. Therefore, we have to evaluate the policies $\{\pi^{\star,j}\}_{j=1}^K$. The policy evaluation and selection algorithm is described in Algorithm 6.

Lemma 15. For Algorithm 6, suppose that the K groups of policies $\{\pi^{1:H,j}\}_{j=1}^K$ and K reward functions $\{(\widehat{r}_i^j)_{i=1}^n\}_{j=1}^K$ satisfy that there exists some $j^\star \in [K]$ such that for any policy $\pi \in \Pi$, $i \in [n]$, we have $|V_{i,1}^{\pi,\mathcal{G}}(\emptyset) - V_{i,1}^{\pi,\widehat{\mathcal{M}}(\pi^{1:H,j^\star})}(\emptyset)| \leq \epsilon$. If $N_2 \geq C\frac{H^2 \log\frac{K^2 n}{\delta_3}}{\epsilon^2}$ for some constant $C > 0$, then with probability at least $1 - \delta_3$, the following event $\mathcal{E}_3$ holds

$$\mathrm{NE/CE/CCE\text{-}gap}(\pi^{\star,\widehat{j}}) \leq \mathrm{NE/CE/CCE\text{-}gap}(\pi^{\star,j^\star}) + 6\epsilon + H\epsilon_e.$$

Proof. For NE/CCE, note that $\pi_i^{\star,j,m} \in \arg\max_{\pi_i} V_{i,1}^{\pi_i \times \pi_{-i}^{\star,j}, \widehat{\mathcal{M}}(\pi^{1:H,m})^m}(\emptyset)$ for $m \in [K]$. By a union bound, with probability at least $1 - \delta_3$, the following event $\mathcal{E}_3$ holds for any $i \in [n], j \in [K], m \in [K]$:

$$|R_i^j - V_{i,1}^{\pi^{\star,j},\mathcal{G}}(\emptyset)| \leq \epsilon,$$
$$|R_i^{j,m} - V_{i,1}^{\pi_i^{\star,j,m} \times \pi_{-i}^{\star,j},\mathcal{G}}(\emptyset)| \leq \epsilon.$$

In the following proof, we will assume the previous event holds. Define $m_{i,j}^\star = \arg\max_m R_i^{j,m}$. Now we will firstly show that $\max_m R_i^{j,m}$ approximates the best response of $\pi_{-i}^{\star,j}$. Note that for any $i \in [n], j \in [K]$:

$$\max_{\pi_i} V_{i,1}^{\pi_i \times \pi_{-i}^{\star,j},\mathcal{G}}(\emptyset) - \max_m R_i^{j,m} \geq \max_{\pi_i} V_{i,1}^{\pi_i \times \pi_{-i}^{\star,j},\mathcal{G}}(\emptyset) - V_{i,1}^{\pi_i^{\star,j,m_{i,j}^\star} \times \pi_{-i}^{\star,j},\mathcal{G}}(\emptyset) - \epsilon \geq -\epsilon.$$

On the other hand,

$$\begin{aligned}
\max_{\pi_i} V_{i,1}^{\pi_i\times\pi_{-i}^{\star,j},\mathcal{G}}(\emptyset) - \max_m R_i^{j,m} &\le \max_{\pi_i} V_{i,1}^{\pi_i\times\pi_{-i}^{\star,j},\mathcal{G}}(\emptyset) - \max_m V_{i,1}^{\pi_i^{\star,j,m}\times\pi_{-i}^{\star,j},\mathcal{G}}(\emptyset) + \epsilon \\
&\le \max_{\pi_i} V_{i,1}^{\pi_i\times\pi_{-i}^{\star,j},\mathcal{G}}(\emptyset) - \max_m V_{i,1}^{\pi_i^{\star,j,m}\times\pi_{-i}^{\star,j},\widehat{\mathcal{M}}(\pi^{1:H,j^\star})}(\emptyset) + 2\epsilon \\
&\le \max_{\pi_i} V_{i,1}^{\pi_i\times\pi_{-i}^{\star,j},\mathcal{G}}(\emptyset) - V_{i,1}^{\pi_i^{\star,j,j^\star}\times\pi_{-i}^{\star,j},\widehat{\mathcal{M}}(\pi^{1:H,j^\star})}(\emptyset) + 2\epsilon \\
&\le \max_{\pi_i} V_{i,1}^{\pi_i\times\pi_{-i}^{\star,j},\mathcal{G}}(\emptyset) - \max_{\pi_i} V_{i,1}^{\pi_i\times\pi_{-i}^{\star,j},\widehat{\mathcal{M}}(\pi^{1:H,j^\star})}(\emptyset) + 2\epsilon + H\epsilon_e \\
&\le 3\epsilon + H\epsilon_e,
\end{aligned}$$

where the second last step comes from Lemma 2 and the last step comes from the fact that the max-operator is non-expansive. Now we are ready to evaluate $\pi^{\star,\widehat{j}}$:

$$\begin{aligned}
\text{NE/CCE-gap}(\pi^{\star,\widehat{j}}) &= \max_i \max_{\pi_i}\left(V_{i,1}^{\pi_i\times\pi_{-i}^{\star,\widehat{j}},\mathcal{G}}(\emptyset) - V_{i,1}^{\pi^{\star,\widehat{j}},\mathcal{G}}(\emptyset)\right) \\
&\le \max_i \max_{\pi_i}\left(V_{i,1}^{\pi_i\times\pi_{-i}^{\star,\widehat{j}},\mathcal{G}}(\emptyset) - R_i^{\widehat{j}}\right) + \epsilon \\
&\le \max_i\left(\max_m R_i^{\widehat{j},m} - R_i^{\widehat{j}}\right) + 4\epsilon + H\epsilon_e.
\end{aligned}$$

Meanwhile for $\pi^{\star,\widehat{j}}$, we have that

$$\begin{aligned}
\text{NE/CCE-gap}(\pi^{\star,j^\star}) &= \max_i \max_{\pi_i}\left(V_{i,1}^{\pi_i\times\pi_{-i}^{\star,j^\star},\mathcal{G}}(\emptyset) - V_{i,1}^{\pi^{\star,j^\star},\mathcal{G}}(\emptyset)\right) \\
&\ge \max_i \max_{\pi_i}\left(V_{i,1}^{\pi_i\times\pi_{-i}^{\star,j^\star},\mathcal{G}}(\emptyset) - R_i^{j^\star}\right) - \epsilon \\
&\ge \max_i\left(\max_m R_i^{j^\star,m} - R_i^{j^\star}\right) - 2\epsilon.
\end{aligned}$$

Recall the definition of $\widehat{j} = \arg\min_j\left(\max_i \max_m(R_i^{j,m} - R_i^j)\right)$, we conclude that $\text{NE/CCE-gap}(\pi^{\star,\widehat{j}}) \le \text{NE-gap}(\pi^{\star,j^\star}) + 6\epsilon + H\epsilon_e$.

For CE, let $\phi_i^{\star,j,m} \in \arg\max_{\phi_i} V_{i,1}^{(\phi_i\diamond\pi_i^{\star,j})\odot\pi_{-i}^{\star,j},\widehat{\mathcal{M}}(\pi^{1:H,m})^m}(\emptyset)$, write $\pi_i^{\star,j,m} := \phi_i^{\star,j,m}\diamond\pi_i^{\star,j}$ for $m\in[K]$. Similarly, by a union bound, with probability at least $1-\delta_3$, the following event $\mathcal{E}_3$ holds for any $i\in[n], j\in[K], m\in[K]$:

$$\begin{aligned}
|R_i^j - V_{i,1}^{\pi^{\star,j},\mathcal{G}}(\emptyset)| &\le \epsilon, \\
|R_i^{j,m} - V_{i,1}^{\pi_i^{\star,j,m}\odot\pi_{-i}^{\star,j},\mathcal{G}}(\emptyset)| &\le \epsilon.
\end{aligned}$$

In the following proof, we will assume the previous event holds. Define $m_{i,j}^\star = \arg\max_m R_i^{j,m}$. Now we will firstly show that $\max_m R_i^{j,m}$ approximates the best strategy modification with respect to $\pi_{-i}^{\star,j}$. Note that for any $i\in[n], j\in[K]$:

$$\max_{\phi_i} V_{i,1}^{(\phi_i\diamond\pi_i^{\star,j})\odot\pi_{-i}^{\star,j},\mathcal{G}}(\emptyset) - \max_m R_i^{j,m} \ge \max_{\phi_i} V_{i,1}^{(\phi_i\diamond\pi_i^{\star,j})\odot\pi_{-i}^{\star,j},\mathcal{G}}(\emptyset) - V_{i,1}^{\pi_i^{\star,j,m_{i,j}^\star}\odot\pi_{-i}^{\star,j},\mathcal{G}}(\emptyset) - \epsilon \ge -\epsilon.$$

On the other hand,

$$\begin{aligned}
\max_{\phi_i} V_{i,1}^{(\phi_i\diamond\pi_i^{\star,j})\odot\pi_{-i}^{\star,j},\mathcal{G}}(\emptyset) - \max_m R_i^{j,m} &\le \max_{\phi_i} V_{i,1}^{(\phi_i\diamond\pi_i^{\star,j})\odot\pi_{-i}^{\star,j},\mathcal{G}}(\emptyset) - \max_m V_{i,1}^{\pi_i^{\star,j,m}\odot\pi_{-i}^{\star,j},\mathcal{G}}(\emptyset) + \epsilon \\
&\le \max_{\phi_i} V_{i,1}^{(\phi_i\diamond\pi_i^{\star,j})\odot\pi_{-i}^{\star,j},\mathcal{G}}(\emptyset) - \max_m V_{i,1}^{\pi_i^{\star,j,m}\odot\pi_{-i}^{\star,j},\widehat{\mathcal{M}}(\pi^{1:H,j^\star})}(\emptyset) + 2\epsilon \\
&\le \max_{\phi_i} V_{i,1}^{(\phi_i\diamond\pi_i^{\star,j})\odot\pi_{-i}^{\star,j},\mathcal{G}}(\emptyset) - V_{i,1}^{\pi_i^{\star,j,j^\star}\odot\pi_{-i}^{\star,j},\widehat{\mathcal{M}}(\pi^{1:H,j^\star})}(\emptyset) + 2\epsilon \\
&\le \max_{\phi_i} V_{i,1}^{(\phi_i\diamond\pi_i^{\star,j})\odot\pi_{-i}^{\star,j},\mathcal{G}}(\emptyset) - \max_{\phi_i} V_{i,1}^{(\phi_i\diamond\pi_i^{\star,j})\odot\pi_{-i}^{\star,j},\widehat{\mathcal{M}}(\pi^{1:H,j^\star})}(\emptyset) + 2\epsilon + H\epsilon_e \\
&\le 3\epsilon + H\epsilon_e,
\end{aligned}$$

where the second last step comes from Lemma 3 and the last step comes from the fact that the max-operator is non-expansive. Now we are ready to evaluate $\pi^{\star,\widehat{j}}$:

$$\begin{aligned}
\text{CE-gap}(\pi^{\star,\widehat{j}}) &= \max_i \max_{\phi_i}\left(V_{i,1}^{(\phi_i\diamond\pi_i^{\star,\widehat{j}})\odot\pi_{-i}^{\star,\widehat{j}},\mathcal{G}}(\emptyset) - V_{i,1}^{\pi^{\star,\widehat{j}},\mathcal{G}}(\emptyset)\right) \\
&\le \max_i \max_{\phi_i}\left(V_{i,1}^{(\phi_i\diamond\pi_i^{\star,\widehat{j}})\odot\pi_{-i}^{\star,\widehat{j}},\mathcal{G}}(\emptyset) - R_i^{\widehat{j}}\right) + \epsilon \\
&\le \max_i\left(\max_m R_i^{\widehat{j},m} - R_i^{\widehat{j}}\right) + 4\epsilon + H\epsilon_e.
\end{aligned}$$

Meanwhile for $\pi^{\star,\widehat{j}}$, we have that

$$\begin{aligned}
\text{CE-gap}(\pi^{\star,j^\star}) &= \max_i \max_{\phi_i}\left(V_{i,1}^{(\phi_i\diamond\pi_i^{\star,j^\star})\odot\pi_{-i}^{\star,j^\star},\mathcal{G}}(\emptyset) - V_{i,1}^{\pi^{\star,j^\star},\mathcal{G}}(\emptyset)\right) \\
&\ge \max_i \max_{\phi_i}\left(V_{i,1}^{(\phi_i\diamond\pi_i^{\star,j^\star})\odot\pi_{-i}^{\star,j^\star},\mathcal{G}}(\emptyset) - R_i^{j^\star}\right) - \epsilon \\
&\ge \max_i\left(\max_m R_i^{j^\star,m} - R_i^{j^\star}\right) - 2\epsilon.
\end{aligned}$$

Recall the definition of $\widehat{j} = \arg\min_j\left(\max_i \max_m (R_i^{j,m} - R_i^j)\right)$, we conclude that $\text{CE-gap}(\pi^{\star,\widehat{j}}) \le \text{CE-gap}(\pi^{\star,j^\star}) + 6\epsilon + H\epsilon_e$. □

We put together the entire learning procedure in Algorithm 7. In the following discussion, we will see the sample complexity of our algorithm instantiated with particular information structures.

One-step delayed information sharing. For this, the information structure has $c_h = \{a_{1:h-1}, o_{2:h-1}\}$, $p_{i,h} = \{o_{i,h}\}$, $z_{h+1} = \{o_h, a_h\}$. Fix $L > 0$, we define the approximate common information as $\widehat{c}_h = \{a_{h-L:h-1}, o_{h-L+1:h-1}\}$. For any $\pi^{1:H}$, it is easy to verify that

$$\mathbb{P}_h^{\widetilde{\mathcal{M}}(\pi^{1:H}),c}(s_h, p_h|\widehat{c}_h) = \mathbb{P}_h^{\pi^h,\mathcal{G}}(s_h, p_h|\widehat{c}_h) = \widetilde{\boldsymbol{b}}_h^{\pi^h}(a_{h-L:h-1}, o_{h-L+1:h-1})(s_h)\mathbb{O}_h(o_h|s_h).$$

Meanwhile, $\widehat{L} = L$. Therefore, we conclude that if $L \geq C\frac{\log(HSO/(\epsilon\gamma))}{\gamma^4}$, by a union bound, with probability at least $1-\delta_1-\delta_2-\delta_3$, it holds that

$$\begin{aligned}
\epsilon_r(\pi^{1:H,j^\star},\widehat{r}) &= \max_h \max_{\pi,\gamma_h} \mathbb{E}^{\mathcal{G}}_{a_{1:h-1},o_{2:h}\sim\pi} |\mathbb{E}^{\mathcal{G}}[r_{i,h+1}(o_{h+1}) \mid c_h,\gamma_h] - \mathbb{E}^{\widetilde{\mathcal{M}}}[\widehat{r}_{i,h+1}(o_{h+1}) \mid \widehat{c}_h,\gamma_h]| \\
&= \max_h \max_{\pi,\gamma_h} \mathbb{E}^{\mathcal{G}}_{a_{1:h-1},o_{2:h}\sim\pi} |\mathbb{E}^{\mathcal{G}}[r_{i,h+1}(o_{h+1}) \mid c_h,\gamma_h] - \mathbb{E}^{\mathcal{G}}[\widehat{r}_{i,h+1}(o_{h+1}) \mid c_h,\gamma_h]| \\
&\quad + \max_h \max_{\pi,\gamma_h} \mathbb{E}^{\mathcal{G}}_{a_{1:h-1},o_{2:h}\sim\pi} |\mathbb{E}^{\mathcal{G}}[\widehat{r}_{i,h+1}(o_{h+1}) \mid c_h,\gamma_h] - \mathbb{E}^{\widetilde{\mathcal{M}}}[\widehat{r}_{i,h+1}(o_{h+1}) \mid \widehat{c}_h,\gamma_h]| \\
&\leq \epsilon + \max_h \max_{\pi} \mathbb{E}^{\mathcal{G}}_{a_{1:h-1},o_{2:h}\sim\pi} \|\boldsymbol{b}_h(a_{1:h-1},o_{2:h-1}) - \widetilde{\boldsymbol{b}}^{\pi^{h,j^\star}}_h(a_{h-L:h-1},o_{h-L+1:h-1})\|_1 \\
&\leq 2\epsilon + \max_h \max_{\pi} \mathbb{1}[h>L]\cdot 6\cdot d^{\pi,\mathcal{G}}_{\mathcal{S},h-L}\left(\mathcal{U}^{\mathcal{G}}_{\phi,h-L}\left(\pi^{h,j^\star}\right)\right).
\end{aligned}$$

$$\begin{aligned}
\epsilon_z(\pi^{1:H,j^\star}) &= \max_h \max_{\pi,\gamma_h} \mathbb{E}^{\mathcal{G}}_{a_{1:h-1},o_{2:h}\sim\pi} \|\mathbb{P}^{\mathcal{G}}_h(\cdot\,|\,c_h,\gamma_h) - \mathbb{P}^{\widetilde{\mathcal{M}},z}_h(\cdot\,|\,c_h,\gamma_h)\|_1 \\
&\leq \max_h \max_{\pi,\gamma_h} \mathbb{E}^{\mathcal{G}}_{a_{1:h-1},o_{2:h}\sim\pi'} \|\boldsymbol{b}_h(a_{1:h-1},o_{2:h-1}) - \widetilde{\boldsymbol{b}}^{\pi^{h,j^\star}}_h(a_{h-L:h-1},o_{h-L+1:h-1})\|_1 \\
&\leq \epsilon + \max_h \max_{\pi} \mathbb{1}[h>L]\cdot 6\cdot d^{\pi,\mathcal{G}}_{\mathcal{S},h-L}\left(\mathcal{U}^{\mathcal{G}}_{\phi,h-L}\left(\pi^{h,j^\star}\right)\right).
\end{aligned}$$

According to the choice $\pi^{1:H,j^\star}$, it holds that

$$\max_h \max_{\pi} \mathbb{1}[h>L]\cdot 6\cdot d^{\pi,\mathcal{G}}_{\mathcal{S},h-L}\left(\mathcal{U}^{\mathcal{G}}_{\phi,h-L}\left(\pi^{h,j^\star}\right)\right) \leq 6\epsilon.$$

Therefore, for any $\alpha,\delta>0$, setting $\epsilon = \frac{\alpha}{200(H+1)^2}$, $\theta_1 = \frac{\alpha}{200(H+1)^2 O}$, $\zeta_2 = \zeta_1^2$, $\theta_2 = \frac{\alpha}{200(H+1)^2 A\max_h P_h}$, $\zeta_1 = \min\{\frac{\alpha\phi}{200(H+1)^2A^{2L}O^L}, \frac{\alpha}{400(H+1)^2A\max_h P_h}\}$, $\phi = \frac{\epsilon\gamma^2}{C^2H^8S^5O^4}$, $\epsilon_e = \frac{\alpha}{200H}$, $\delta_1=\delta_2=\delta_3=\frac{\delta}{3}$, $\widehat{\mathcal{M}}(\pi^{1:H,j^\star})$ is an (ϵ_r,ϵ_z)-expected-approximate common information model of $\mathcal{G}$, where $\epsilon_r,\epsilon_z \leq \frac{14\alpha}{200(H+1)^2}$. This leads to that $\pi^{\star,j^\star}$ is a $\frac{15\alpha}{200}$-NE/CE/CCE, and $|V^{\pi,\mathcal{G}}_{i,1}(\emptyset) - V^{\pi,\widehat{\mathcal{M}}(\pi^{1:H,j^\star})}_{i,1}(\emptyset)| \leq \frac{15\alpha}{200}$ for any policy $\pi\in\Pi$ by Lemma 4. By Lemma 15, NE/CE/CCE-gap$(\pi^{\star,\widehat{j}}) \leq$ NE/CE/CCE-gap$(\pi^{\star,j^\star}) + \frac{91\alpha}{200} \leq \alpha$. Finally, we are ready to analyze the computation and sample complexity of our algorithm.

Theorem 16. Let $\alpha,\delta,\gamma>0$. Algorithm 7 given a γ-observable POSG of one-step delayed information sharing structure has time and sample complexity bounded by $(AO)^{C\gamma^{-4}\log\frac{SHO}{\gamma\alpha}}\log\frac{1}{\delta}$ for some universal constant $C>0$ outputting an α-NE/CE/CCE with probability at least $1-\delta$.

Proof. Recall that $\widehat{C}_h \leq (OA)^L$, $P_h \leq O$, $N_0 = \max\{\frac{C(\max_h P_h + \log\frac{4H\max_h \widehat{C}_h}{\delta_1})}{\zeta_1\theta_1^2}, \frac{CA(O+\log\frac{4H\max_h(\widehat{C}_h P_h A)}{\delta_1})}{\zeta_2\theta_2^2}\}$, $N_1 = (OA)^L\log(\frac{1}{\delta_2})$, and $N_2 = C\frac{H^2\log\frac{K^2_n}{\delta_3}}{\epsilon^2}$ for some constant $C>0$. The total number of samples used is $KN_0+N_1+(K+nK^2)N_2$. Substituting the choices of parameters into N_0, N_1, and N_2, we proved the sample complexity. Furthermore, for time complexity analysis, since our algorithm only calls the `BaseCAMP` and our planning algorithm polynomial number of times, time complexity is also bounded by $(OA)^{C\gamma^{-4}\log\frac{SHO}{\gamma\alpha}}\log\frac{1}{\delta}$. □

State controlled by one controller with asymmetric delay sharing. The information structure is given as $c_h = \{o_{1,2:h}, o_{2,2:h-d}, a_{1,1:h-1}\}$, $p_{1,h} = \emptyset$, $p_{2,h} = \{o_{2,h-d+1:h}\}$. Fix some $L\geq 0$, the approximate common information is constructed as $\widehat{c}_h := \{o_{1,h-d-L+1:h}, o_{2,h-d-L+1:h-d}, a_{1,h-d-L:h-1}\}$. Then for any given policy $\pi^{1:H}$, following exactly the same derivation as in (E.14), it holds that

$$\mathbb{P}^{\widetilde{\mathcal{M}}(\pi^{1:H}),c}_h(s_h,p_h|\widehat{c}_h) = \mathbb{P}^{\pi^h,\mathcal{G}}_h(s_h,p_h|\widehat{c}_h) = \sum_{s_{h-d}} \mathbb{P}^{\mathcal{G}}(s_h,p_h|s_{h-d},f_a,f_o) F^{P(\cdot|\cdot,f_a)}(\widetilde{\boldsymbol{b}}^{\pi^h}_{h-d}(a_{1:h-d-1},o_{2:h-d});f_o)(s_{h-d}).$$

Meanwhile, $\widehat{L} = L + d$. Therefore, we conclude that if $L \geq C\frac{\log(HSO/(\epsilon\gamma))}{\gamma^4}$, by a union bound, with probability at least $1-\delta_1-\delta_2-\delta_3$:

$$\begin{aligned}
\epsilon_r(\pi^{1:H,j^\star}, \widehat{r}) &= \max_h \max_{\pi,\gamma_h} \mathbb{E}^{\mathcal{G}}_{a_{1:h-1},o_{2:h}\sim\pi} |\mathbb{E}^{\mathcal{G}}[r_{i,h+1}(o_{h+1}) \mid c_h, \gamma_h] - \mathbb{E}^{\widetilde{\mathcal{M}}}[\widehat{r}_{i,h+1}(o_{h+1}) \mid \widehat{c}_h, \gamma_h]| \\
&= \max_h \max_{\pi,\gamma_h} \mathbb{E}^{\mathcal{G}}_{a_{1:h-1},o_{2:h}\sim\pi} |\mathbb{E}^{\mathcal{G}}[r_{i,h+1}(o_{h+1}) \mid c_h, \gamma_h] - \mathbb{E}^{\mathcal{G}}[\widehat{r}_{i,h+1}(o_{h+1}) \mid c_h, \gamma_h]| \\
&\quad + \max_h \max_{\pi,\gamma_h} \mathbb{E}^{\mathcal{G}}_{a_{1:h-1},o_{2:h}\sim\pi} |\mathbb{E}^{\mathcal{G}}[\widehat{r}_{i,h+1}(o_{h+1}) \mid c_h, \gamma_h] - \mathbb{E}^{\widetilde{\mathcal{M}}}[\widehat{r}_{i,h+1}(o_{h+1}) \mid \widehat{c}_h, \gamma_h]| \\
&\leq \epsilon + \max_h \max_{\pi} \mathbb{E}^{\mathcal{G}}_{a_{1:h-1},o_{2:h}\sim\pi} \|F^{P(\cdot|\cdot,f_a)}(\boldsymbol{b}_{h-d}(a_{1:h-d-1}, o_{2:h-d}); f_o) - F^{P(\cdot|\cdot,f_a)}(\widetilde{\boldsymbol{b}}^{\pi^{h,j^\star}}_{h-d}(a_{h-d-L:h-d-1}, o_{h-d-L+1:h-d}); f_o)\|_1 \\
&\leq 2\epsilon + \max_h \max_{\pi} \mathbb{1}[h > \widehat{L}] \cdot 6 \cdot d^{\pi,\mathcal{G}}_{\mathcal{S},h-\widehat{L}}\left(\mathcal{U}^{\mathcal{G}}_{\phi,h-\widehat{L}}\left(\pi^{h,j^\star}\right)\right).
\end{aligned}$$

$$\begin{aligned}
\epsilon_z(\pi^{1:H,j^\star}) &= \max_h \max_{\pi,\gamma_h} \mathbb{E}^{\mathcal{G}}_{a_{1:h-1},o_{2:h}\sim\pi} \|\mathbb{P}^{\mathcal{G}}_h(\cdot|c_h,\gamma_h) - \mathbb{P}^{\widetilde{\mathcal{M}},z}_h(\cdot|c_h,\gamma_h)\|_1 \\
&\leq \max_h \max_{\pi,\gamma_h} \mathbb{E}^{\mathcal{G}}_{a_{1:h-1},o_{2:h}\sim\pi'} \|F^{P(\cdot|\cdot,f_a)}(\boldsymbol{b}_{h-d}(a_{1:h-d-1}, o_{2:h-d}); f_o) - F^{P(\cdot|\cdot,f_a)}(\widetilde{\boldsymbol{b}}^{\pi^{h,j^\star}}_{h-d}(a_{h-d-L:h-d-1}, o_{h-d-L+1:h-d}); f_o)\|_1 \\
&\leq \epsilon + \max_h \max_{\pi} \mathbb{1}[h > \widehat{L}] \cdot 6 \cdot d^{\pi,\mathcal{G}}_{\mathcal{S},h-\widehat{L}}\left(\mathcal{U}^{\mathcal{G}}_{\phi,h-\widehat{L}}\left(\pi^{h,j^\star}\right)\right).
\end{aligned}$$

According to the choice $\pi^{1:H,j^\star}$, it holds that

$$\max_h \max_{\pi} \mathbb{1}[h > \widehat{L}] \cdot 6 \cdot d^{\pi,\mathcal{G}}_{\mathcal{S},h-\widehat{L}}\left(\mathcal{U}^{\mathcal{G}}_{\phi,h-\widehat{L}}\left(\pi^{h,j^\star}\right)\right) \leq 6\epsilon.$$

Therefore, for any $\alpha, \delta > 0$, setting $\epsilon = \frac{\alpha}{200(H+1)^2}$, $\theta_1 = \frac{\alpha}{200(H+1)^2 O}$, $\zeta_2 = \zeta_1^2$, $\theta_2 = \frac{\alpha}{200(H+1)^2 A \max_h P_h}$, $\zeta_1 = \min\{\frac{\alpha\phi}{200(H+1)^2 A^{2(L+d)} O^{L+d}}, \frac{\alpha}{400(H+1)^2 A \max_h P_h}\}$, $\phi = \frac{\epsilon\gamma^2}{C^2 H^8 S^5 O^4}$, $\epsilon_e = \frac{\alpha}{200H}$, $\delta_1 = \delta_2 = \delta_3 = \frac{\delta}{3}$, $\widehat{\mathcal{M}}(\pi^{1:H,j^\star})$ is an (ϵ_r, ϵ_z)-expected-approximate common information model of $\mathcal{G}$, where $\epsilon_r, \epsilon_z \leq \frac{14\alpha}{200(H+1)^2}$. This leads to that $\pi^{\star,j^\star}$ is a $\frac{15\alpha}{200}$-NE/CE/CCE, and $|V^{\pi,\mathcal{G}}_{i,1}(\emptyset) - V^{\pi,\widehat{\mathcal{M}}(\pi^{1:H,j^\star})}_{i,1}(\emptyset)| \leq \frac{15\alpha}{200}$ for any policy $\pi \in \Pi$ by Lemma 4. By Lemma 15, NE/CE/CCE-gap$(\pi^{\star,\widehat{j}}) \leq$ NE/CE/CCE-gap$(\pi^{\star,j^\star}) + \frac{91\alpha}{200} \leq \alpha$. Finally, we are ready to analyze the computation and sample complexity of our algorithm.

Theorem 17. Let $\alpha, \delta, \gamma > 0$. Algorithm 7 given a γ-observable POSG of state controlled by one controller with asymmetric delay sharing has computation and sample complexity bounded by $(OA)^{C(\gamma^{-4}\log\frac{SHO}{\gamma\alpha}+d)}\log\frac{1}{\delta}$ for some universal constant $C > 0$ outputting an α-NE/CE/CCE with probability at least $1-\delta$.

Proof. Recall that $\widehat{C}_h \leq (AO)^L$, $P_h \leq (AO)^d$, $N_0 = \max\{\frac{C(\max_h P_h + \log\frac{4H\max_h \widehat{C}_h}{\delta_1})}{\zeta_1\theta_1^2}, \frac{CA(O+\log\frac{4H\max_h(\widehat{C}_h P_h)A}{\delta_1})}{\zeta_2\theta_2^2}\}$, $N_1 = (OA)^{\widehat{L}}\log(\frac{1}{\delta_2})$, and $N_2 = C\frac{H^2\log\frac{K^2 n}{\delta_3}}{\epsilon^2}$ for some constant $C > 0$. The total number of samples used is $KN_0 + N_1 + (K+nK^2)N_2$. Substituting the choices of parameters into N_0, N_1, and N_2, we proved the sample complexity. Furthermore, for time complexity analysis, since our algorithm only calls the `BaseCAMP` and our planning algorithm polynomial number of times, time complexity is also bounded by $(OA)^{C(\gamma^{-4}\log\frac{SHO}{\gamma\alpha}+d)}\log\frac{1}{\delta}$. □

Information sharing with one-directional-one-step delay. For this case, we have $c_h = \{o_{1,2:h}, o_{2,2:h-1}, a_{1:h-1}\}$, $p_{1,h} = \emptyset$, $p_{2,h} = \{o_{2,h}\}$, and $z_{h+1} = \{o_{1,h+1}, o_{2,h}, a_h\}$. Fix $L > 0$, we construct the approximate common information as $\widehat{c}_h = \{o_{1,h-L+1:h}, o_{2,h-L+1:h-1}, a_{h-L:h-1}\}$. For any $\pi^{1:H}$, it is easy to verify that

$$\mathbb{P}^{\widetilde{\mathcal{M}}(\pi^{1:H}),c}_h(s_h, p_h|\widehat{c}_h) = \mathbb{P}^{\pi^h,\mathcal{G}}_h(s_h, p_h|\widehat{c}_h) = \widetilde{\boldsymbol{b}}^{\pi^h}_h(o_{1,h-L:h}, o_{2,h-L:h-1}, a_{h-L:h-1})(s_h)\mathbb{P}_h(o_{2,h}|s_h, o_{1,h}),$$

where $\mathbb{P}_h(o_{2,h}|s_h, o_{1,h}) = \frac{\mathbb{O}_h(o_{1,h},o_{2,h}|s_h)}{\sum_{o'_{2,h}} \mathbb{O}_h(o_{1,h},o'_{2,h}|s_h)}$. Furthermore, $\widehat{L} = L$. Therefore, we conclude that if $L \geq C\frac{\log(HSO/(\epsilon\gamma))}{\gamma^4}$, by a union bound, with probability at least $1-\delta_1-\delta_2-\delta_3$:

$$\begin{aligned}
\epsilon_r(\pi^{1:H,j^\star}, \widehat{r}) &= \max_h \max_{\pi,\gamma_h} \mathbb{E}^{\mathcal{G}}_{a_{1:h-1},o_{2:h}\sim\pi} |\mathbb{E}^{\mathcal{G}}[r_{i,h+1}(o_{h+1}) \mid c_h, \gamma_h] - \mathbb{E}^{\widetilde{\mathcal{M}}}[\widehat{r}_{i,h+1}(o_{h+1}) \mid \widehat{c}_h, \gamma_h]| \\
&= \max_h \max_{\pi,\gamma_h} \mathbb{E}^{\mathcal{G}}_{a_{1:h-1},o_{2:h}\sim\pi} |\mathbb{E}^{\mathcal{G}}[r_{i,h+1}(o_{h+1}) \mid c_h, \gamma_h] - \mathbb{E}^{\mathcal{G}}[\widehat{r}_{i,h+1}(o_{h+1}) \mid c_h, \gamma_h]| \\
&\quad + \max_h \max_{\pi,\gamma_h} \mathbb{E}^{\mathcal{G}}_{a_{1:h-1},o_{2:h}\sim\pi} |\mathbb{E}^{\mathcal{G}}[\widehat{r}_{i,h+1}(o_{h+1}) \mid c_h, \gamma_h] - \mathbb{E}^{\widetilde{\mathcal{M}}}[\widehat{r}_{i,h+1}(o_{h+1}) \mid \widehat{c}_h, \gamma_h]| \\
&\leq \epsilon + \max_h \max_{\pi} \mathbb{E}^{\mathcal{G}}_{a_{1:h-1},o_{2:h}\sim\pi} \|\boldsymbol{b}_h(a_{1:h-1}, o_{2:h-1}, o_{1,h}) - \widetilde{\boldsymbol{b}}^{\pi^{h,j^\star}}_h(a_{h-L:h-1}, o_{h-L+1:h-1}, o_{1,h})\|_1 \\
&\leq 2\epsilon + \max_h \max_{\pi} \mathbb{1}[h > L] \cdot 6 \cdot d^{\pi,\mathcal{G}}_{\mathcal{S},h-L}\left(\mathcal{U}^{\mathcal{G}}_{\phi,h-L}\left(\pi^{h,j^\star}\right)\right).
\end{aligned}$$

$$\begin{aligned}
\epsilon_z(\pi^{1:H,j^\star}) &= \max_h \max_{\pi,\gamma_h} \mathbb{E}^{\mathcal{G}}_{a_{1:h-1},o_{2:h}\sim\pi} \|\mathbb{P}^{\mathcal{G}}_h(\cdot \mid c_h, \gamma_h) - \mathbb{P}^{\widetilde{\mathcal{M}},z}_h(\cdot \mid c_h, \gamma_h)\|_1 \\
&\leq \max_h \max_{\pi,\gamma_h} \mathbb{E}^{\mathcal{G}}_{a_{1:h-1},o_{2:h}\sim\pi'} \|\boldsymbol{b}_h(a_{1:h-1}, o_{2:h-1}, o_{1,h}) - \widetilde{\boldsymbol{b}}^{\pi^{h,j^\star}}_h(a_{h-L:h-1}, o_{h-L+1:h-1}, o_{1,h})\|_1 \\
&\leq \epsilon + \max_h \max_{\pi} \mathbb{1}[h > L] \cdot 6 \cdot d^{\pi,\mathcal{G}}_{\mathcal{S},h-L}\left(\mathcal{U}^{\mathcal{G}}_{\phi,h-L}\left(\pi^{h,j^\star}\right)\right).
\end{aligned}$$

According to the choice $\pi^{1:H,j^\star}$, it holds that

$$\max_h \max_{\pi} \mathbb{1}[h > L] \cdot 6 \cdot d^{\pi,\mathcal{G}}_{\mathcal{S},h-L}\left(\mathcal{U}^{\mathcal{G}}_{\phi,h-L}\left(\pi^{h,j^\star}\right)\right) \leq 6\epsilon.$$

Therefore, for any $\alpha, \delta > 0$, setting $\epsilon = \frac{\alpha}{200(H+1)^2}$, $\theta_1 = \frac{\alpha}{200(H+1)^2 O}$, $\zeta_2 = \zeta_1^2$, $\theta_2 = \frac{\alpha}{200(H+1)^2 A \max_h P_h}$, $\zeta_1 = \min\{\frac{\alpha\phi}{200(H+1)^2 A^{2L} O^L}, \frac{\alpha}{400(H+1)^2 A \max_h P_h}\}$, $\phi = \frac{\epsilon\gamma^2}{C^2 H^8 S^5 O^4}$, $\epsilon_e = \frac{\alpha}{200H}$, $\delta_1 = \delta_2 = \delta_3 = \frac{\delta}{3}$, $\widehat{\mathcal{M}}(\pi^{1:H,j^\star})$ is an (ϵ_r, ϵ_z)-expected-approximate common information model of $\mathcal{G}$, where $\epsilon_r, \epsilon_z \leq \frac{14\alpha}{200(H+1)^2}$. This leads to that $\pi^{\star,j^\star}$ is a $\frac{15\alpha}{200}$-NE/CE/CCE, and $|V^{\pi,\mathcal{G}}_{i,1}(\emptyset) - V^{\pi,\widehat{\mathcal{M}}(\pi^{1:H,j^\star})}_{i,1}(\emptyset)| \leq \frac{15\alpha}{200}$ for any policy $\pi \in \Pi$ by Lemma 4. By Lemma 15, NE/CE/CCE-gap$(\pi^{\star,\widehat{j}}) \leq$ NE/CE/CCE-gap$(\pi^{\star,j^\star}) + \frac{91\alpha}{200} \leq \alpha$. Finally, we are ready to analyze the computation and sample complexity of our algorithm.

Theorem 18. Let $\alpha, \delta, \gamma > 0$. Algorithm 7 given a γ-observable POSG of one-directional-one-step delayed information sharing structure has time and sample complexity bounded by $(AO)^{C\gamma^{-4}\log\frac{SHO}{\gamma\alpha}} \log\frac{1}{\delta}$ for some universal constant $C > 0$ outputting an α-NE/CE/CCE with probability at least $1-\delta$.

Proof. Recall that $\widehat{C}_h \leq (OA)^L$, $P_h \leq O$, $N_0 = \max\{\frac{C(\max_h P_h + \log\frac{4H\max_h \widehat{C}_h}{\delta_1})}{\zeta_1\theta_1^2}, \frac{CA(O+\log\frac{4H\max_h(\widehat{C}_h P_h A)}{\delta_1})}{\zeta_2\theta_2^2}\}$, $N_1 = (OA)^L \log(\frac{1}{\delta_2})$, and $N_2 = C\frac{H^2\log\frac{K^2 n}{\delta_3}}{\epsilon^2}$ for some constant $C > 0$. The total number of samples used is $KN_0 + N_1 + (K + nK^2)N_2$. Substituting the choices of parameters into N_0, N_1, and N_2, we proved the sample complexity. Furthermore, for time complexity analysis, since our algorithm only calls the `BaseCAMP` and our planning algorithm polynomial number of times, time complexity is also bounded by $(OA)^{C\gamma^{-4}\log\frac{SHO}{\gamma\alpha}} \log\frac{1}{\delta}$. □

Uncontrolled state process with delayed sharing. The information structure satisfies $c_h = \{o_{2:h-d}\}$, $p_{i,h} = \{o_{i,h-d+1:h}\}$, and $z_{h+1} = \{o_{h-d+1}\}$. Fix a $L \geq 0$, the approximate common information is $\widehat{c}_h = \{o_{h-d-L+1:h-d}\}$. For any policy $\pi^{1:H}$, it is easy to verify that

$$\mathbb{P}^{\widetilde{\mathcal{M}}(\pi^{1:H}),c}_h(s_h, p_h|\widehat{c}_h) = \mathbb{P}^{\pi^h,\mathcal{G}}_h(s_h, p_h|\widehat{c}_h) = \sum_{s_{h-d}} \widetilde{\boldsymbol{b}}^{\pi^h}_{h-d}(o_{h-d-L+1:h-d})(s_{h-d})\mathbb{P}(s_h, o_{h-d+1:h}|s_{h-d}).$$

Furthermore, $\widehat{L} = L + d$. Therefore, we conclude that if $L \geq C\frac{\log(HSO/(\epsilon\gamma))}{\gamma^4}$, by a union bound, with probability at least $1-\delta_1-\delta_2-\delta_3$:

$$\begin{aligned}
\epsilon_r(\pi^{1:H,j^\star}, \widehat{r}) &= \max_h \max_{\pi,\gamma_h} \mathbb{E}^{\mathcal{G}}_{a_{1:h-1}, o_{2:h}\sim\pi} |\mathbb{E}^{\mathcal{G}}[r_{i,h+1}(o_{h+1}) \mid c_h, \gamma_h] - \mathbb{E}^{\widetilde{\mathcal{M}}}[\widehat{r}_{i,h+1}(o_{h+1}) \mid \widehat{c}_h, \gamma_h]| \\
&= \max_h \max_{\pi,\gamma_h} \mathbb{E}^{\mathcal{G}}_{a_{1:h-1}, o_{2:h}\sim\pi} |\mathbb{E}^{\mathcal{G}}[r_{i,h+1}(o_{h+1}) \mid c_h, \gamma_h] - \mathbb{E}^{\mathcal{G}}[\widehat{r}_{i,h+1}(o_{h+1}) \mid c_h, \gamma_h]| \\
&\quad + \max_h \max_{\pi,\gamma_h} \mathbb{E}^{\mathcal{G}}_{a_{1:h-1}, o_{2:h}\sim\pi} |\mathbb{E}^{\mathcal{G}}[\widehat{r}_{i,h+1}(o_{h+1}) \mid c_h, \gamma_h] - \mathbb{E}^{\widetilde{\mathcal{M}}}[\widehat{r}_{i,h+1}(o_{h+1}) \mid \widehat{c}_h, \gamma_h]| \\
&\leq \epsilon + \max_h \max_{\pi} \mathbb{E}^{\mathcal{G}}_{a_{1:h-1}, o_{2:h}\sim\pi} \|\boldsymbol{b}_{h-d}(o_{2:h-d}) - \widetilde{\boldsymbol{b}}^{\pi^{h,j^\star}}_{h-d}(o_{h-d-L+1:h-d})\|_1 \\
&\leq 2\epsilon + \max_h \max_{\pi} \mathbb{1}[h > \widehat{L}] \cdot 6 \cdot d^{\pi,\mathcal{G}}_{\mathcal{S},h-\widehat{L}}\left(\mathcal{U}^{\mathcal{G}}_{\phi,h-\widehat{L}}\left(\pi^{h,j^\star}\right)\right).
\end{aligned}$$

$$\begin{aligned}
\epsilon_z(\pi^{1:H,j^\star}) &= \max_h \max_{\pi,\gamma_h} \mathbb{E}^{\mathcal{G}}_{a_{1:h-1}, o_{2:h}\sim\pi} \|\mathbb{P}^{\mathcal{G}}_h(\cdot \mid c_h, \gamma_h) - \mathbb{P}^{\widetilde{\mathcal{M}},z}_h(\cdot \mid c_h, \gamma_h)\|_1 \\
&\leq \max_h \max_{\pi,\gamma_h} \mathbb{E}^{\mathcal{G}}_{a_{1:h-1}, o_{2:h}\sim\pi'} \|\boldsymbol{b}_{h-d}(o_{2:h-d}) - \widetilde{\boldsymbol{b}}^{\pi^{h,j^\star}}_{h-d}(o_{h-d-L+1:h-d})\|_1 \\
&\leq \epsilon + \max_h \max_{\pi} \mathbb{1}[h > \widehat{L}] \cdot 6 \cdot d^{\pi,\mathcal{G}}_{\mathcal{S},h-\widehat{L}}\left(\mathcal{U}^{\mathcal{G}}_{\phi,h-\widehat{L}}\left(\pi^{h,j^\star}\right)\right).
\end{aligned}$$

According to the choice $\pi^{1:H,j^\star}$, it holds that

$$\max_h \max_{\pi} \mathbb{1}[h > \widehat{L}] \cdot 6 \cdot d^{\pi,\mathcal{G}}_{\mathcal{S},h-\widehat{L}}\left(\mathcal{U}^{\mathcal{G}}_{\phi,h-\widehat{L}}\left(\pi^{h,j^\star}\right)\right) \leq 6\epsilon.$$

Therefore, for any $\alpha, \delta > 0$, setting $\epsilon = \frac{\alpha}{200(H+1)^2}$, $\theta_1 = \frac{\alpha}{200(H+1)^2 O}$, $\zeta_2 = \zeta_1^2$, $\theta_2 = \frac{\alpha}{200(H+1)^2 A \max_h P_h}$, $\zeta_1 = \min\{\frac{\alpha\phi}{200(H+1)^2 A^{2(L+d)} O^{L+d}}, \frac{\alpha}{400(H+1)^2 A \max_h P_h}\}$, $\phi = \frac{\epsilon\gamma^2}{C^2 H^8 S^5 O^4}$, $\epsilon_e = \frac{\alpha}{200H}$, $\delta_1 = \delta_2 = \delta_3 = \frac{\delta}{3}$, $\widehat{\mathcal{M}}(\pi^{1:H,j^\star})$ is an (ϵ_r, ϵ_z)-expected-approximate common information model of $\mathcal{G}$, where $\epsilon_r, \epsilon_z \leq \frac{14\alpha}{200(H+1)^2}$. This leads to that $\pi^{\star,j^\star}$ is a $\frac{15\alpha}{200}$-NE/CE/CCE, and $|V^{\pi,\mathcal{G}}_{i,1}(\emptyset) - V^{\pi,\widehat{\mathcal{M}}(\pi^{1:H,j^\star})}_{i,1}(\emptyset)| \leq \frac{15\alpha}{200}$ for any policy π by Lemma 4. By Lemma 15, NE/CE/CCE-gap$(\pi^{\star,\widehat{j}}) \leq$ NE/CE/CCE-gap$(\pi^{\star,j^\star}) + \frac{91\alpha}{200} \leq \alpha$. Finally, we are ready to analyze the computation and sample complexity of our algorithm.

Theorem 19. Let $\alpha, \delta, \gamma > 0$. Algorithm 7 given a γ-observable POSG of uncontrolled state process and delayed information sharing structure has time and sample complexity bounded by $(OA)^{C(\gamma^{-4}\log\frac{SHO}{\gamma\alpha}+d)}\log\frac{1}{\delta}$ for some universal constant $C > 0$ outputting an α-NE/CE/CCE with probability at least $1-\delta$.

Proof. Recall that $\widehat{C}_h \leq O^L$, $P_h \leq O^d$, $N_0 = \max\{\frac{C(\max_h P_h + \log\frac{4H\max_h \widehat{C}_h}{\delta_1})}{\zeta_1\theta_1^2}, \frac{CA(O+\log\frac{4H\max_h(\widehat{C}_h P_h)A}{\delta_1})}{\zeta_2\theta_2^2}\}$, $N_1 = (OA)^{\widehat{L}}\log(\frac{1}{\delta_2})$, and $N_2 = C\frac{H^2\log\frac{K^2 n}{\delta_3}}{\epsilon^2}$ for some constant $C > 0$. The total number of samples used is $KN_0 + N_1 + (K + nK^2)N_2$. Substituting the choices of parameters into N_0, N_1, and N_2, we proved the sample complexity. Furthermore, for time complexity analysis, since our algorithm only calls the `BaseCAMP` and our planning algorithm polynomial number of times, time complexity is also bounded by $(OA)^{C(\gamma^{-4}\log\frac{SHO}{\gamma\alpha})}\log\frac{1}{\delta}$. □

Symmetric information game. For symmetric information game, $c_h = \{o_{1:h}, a_{1:h-1}\}$, $p_{i,h} = \emptyset$, and $z_{h+1} = \{a_h, o_{h+1}\}$. Fix $L \geq 0$, we construct the approximate common information as $\widehat{c}_h = \{o_{h-L+1:h}, a_{h-L:h-1}\}$. For any $\pi^{1:H}$, it is easy to verify that

$$\mathbb{P}^{\widetilde{\mathcal{M}}(\pi^{1:H}),c}_h(s_h, p_h \mid \widehat{c}_h) = \mathbb{P}^{\pi^h,\mathcal{G}}_h(s_h, p_h \mid \widehat{c}_h) = \widetilde{\boldsymbol{b}}^{\pi^h}_h(a_{h-L:h-1}, o_{h-L+1:h})(s_h).$$

Meanwhile, $\widehat{L} = L$. Therefore, we conclude that if $L \geq C\frac{\log(HSO/(\epsilon\gamma))}{\gamma^4}$, by a union bound, with probability at least $1-\delta_1-\delta_2-\delta_3$:

$$\begin{aligned}
\epsilon_r(\pi^{1:H,j^\star}, \widehat{r}) &= \max_h \max_{\pi,\gamma_h} \mathbb{E}^{\mathcal{G}}_{a_{1:h-1},o_{2:h}\sim\pi} |\mathbb{E}^{\mathcal{G}}[r_{i,h+1}(o_{h+1}) \mid c_h, \gamma_h] - \mathbb{E}^{\widetilde{\mathcal{M}}}[\widehat{r}_{i,h+1}(o_{h+1}) \mid \widehat{c}_h, \gamma_h]| \\
&= \max_h \max_{\pi,\gamma_h} \mathbb{E}^{\mathcal{G}}_{a_{1:h-1},o_{2:h}\sim\pi} |\mathbb{E}^{\mathcal{G}}[r_{i,h+1}(o_{h+1}) \mid c_h, \gamma_h] - \mathbb{E}^{\mathcal{G}}[\widehat{r}_{i,h+1}(o_{h+1}) \mid c_h, \gamma_h]| \\
&\quad + \max_h \max_{\pi,\gamma_h} \mathbb{E}^{\mathcal{G}}_{a_{1:h-1},o_{2:h}\sim\pi} |\mathbb{E}^{\mathcal{G}}[\widehat{r}_{i,h+1}(o_{h+1}) \mid c_h, \gamma_h] - \mathbb{E}^{\widetilde{\mathcal{M}}}[\widehat{r}_{i,h+1}(o_{h+1}) \mid \widehat{c}_h, \gamma_h]| \\
&\leq \epsilon + \max_h \max_{\pi} \mathbb{E}^{\mathcal{G}}_{a_{1:h-1},o_{2:h}\sim\pi} \|\boldsymbol{b}_h(a_{1:h-1}, o_{2:h}) - \widetilde{\boldsymbol{b}}_h^{\pi^{h,j^\star}}(a_{h-L:h-1}, o_{h-L+1:h})\|_1 \\
&\leq 2\epsilon + \max_h \max_{\pi} \mathbb{1}[h > L] \cdot 6 \cdot d^{\pi,\mathcal{G}}_{\mathcal{S},h-L}\left(\mathcal{U}^{\mathcal{G}}_{\phi,h-L}\left(\pi^{h,j^\star}\right)\right).
\end{aligned}$$

$$\begin{aligned}
\epsilon_z(\pi^{1:H,j^\star}) &= \max_h \max_{\pi,\gamma_h} \mathbb{E}^{\mathcal{G}}_{a_{1:h-1},o_{2:h}\sim\pi} \|\mathbb{P}^{\mathcal{G}}_h(\cdot \mid c_h, \gamma_h) - \mathbb{P}^{\widetilde{\mathcal{M}},z}_h(\cdot \mid c_h, \gamma_h)\|_1 \\
&\leq \max_h \max_{\pi,\gamma_h} \mathbb{E}^{\mathcal{G}}_{a_{1:h-1},o_{2:h}\sim\pi'} \|\boldsymbol{b}_h(a_{1:h-1}, o_{2:h}) - \widetilde{\boldsymbol{b}}_h^{\pi^{h,j^\star}}(a_{h-L:h-1}, o_{h-L+1:h})\|_1 \\
&\leq \epsilon + \max_h \max_{\pi} \mathbb{1}[h > L] \cdot 6 \cdot d^{\pi,\mathcal{G}}_{\mathcal{S},h-L}\left(\mathcal{U}^{\mathcal{G}}_{\phi,h-L}\left(\pi^{h,j^\star}\right)\right).
\end{aligned}$$

According to the choice $\pi^{1:H,j^\star}$, it holds that

$$\max_h \max_{\pi} \mathbb{1}[h > L] \cdot 6 \cdot d^{\pi,\mathcal{G}}_{\mathcal{S},h-L}\left(\mathcal{U}^{\mathcal{G}}_{\phi,h-L}\left(\pi^{h,j^\star}\right)\right) \leq 6\epsilon.$$

Therefore, for any $\alpha, \delta > 0$, setting $\epsilon = \frac{\alpha}{200(H+1)^2}$, $\theta_1 = \frac{\alpha}{200(H+1)^2 O}$, $\zeta_2 = \zeta_1^2$, $\theta_2 = \frac{\alpha}{200(H+1)^2 A \max_h P_h}$, $\zeta_1 = \min\{\frac{\alpha\phi}{200(H+1)^2 A^{2L} O^L}, \frac{\alpha}{400(H+1)^2 A \max_h P_h}\}$, $\phi = \frac{\epsilon\gamma^2}{C^2 H^8 S^5 O^4}$, $\epsilon_e = \frac{\alpha}{200H}$, $\delta_1 = \delta_2 = \delta_3 = \frac{\delta}{3}$, $\widehat{\mathcal{M}}(\pi^{1:H,j^\star})$ is an (ϵ_r, ϵ_z)-expected-approximate common information model of $\mathcal{G}$, where $\epsilon_r, \epsilon_z \leq \frac{14\alpha}{200(H+1)^2}$. This leads to that $\pi^{\star,j^\star}$ is a $\frac{15\alpha}{200}$-NE/CE/CCE, and $|V^{\pi,\mathcal{G}}_{i,1}(\emptyset) - V^{\pi,\widehat{\mathcal{M}}(\pi^{1:H,j^\star})}_{i,1}(\emptyset)| \leq \frac{15\alpha}{200}$ for any policy $\pi \in \Pi$ by Lemma 4. By Lemma 15, NE/CE/CCE-gap$(\pi^{\star,\widehat{j}}) \leq$ NE/CE/CCE-gap$(\pi^{\star,j^\star}) + \frac{91\alpha}{200} \leq \alpha$. Finally, we are ready to analyze the computation and sample complexity of our algorithm.

Theorem 20. Let $\alpha, \delta, \gamma > 0$. Algorithm 7 given a γ-observable POSG of symmetric information sharing structure has time and sample complexity bounded by $(AO)^{C\gamma^{-4}\log\frac{SHO}{\gamma\alpha}}\log\frac{1}{\delta}$ for some universal constant $C > 0$ outputting an α-NE/CE/CCE with probability at least $1-\delta$.

Proof. Recall that $\widehat{C}_h \leq (OA)^L$, $P_h = 1$, $N_0 = \max\{\frac{C(\max_h P_h + \log\frac{4H\max_h \widehat{C}_h}{\delta_1})}{\zeta_1\theta_1^2}, \frac{CA(O+\log\frac{4H\max_h(\widehat{C}_h P_h)A}{\delta_1})}{\zeta_2\theta_2^2}\}$, $N_1 = (OA)^L\log(\frac{1}{\delta_2})$, and $N_2 = C\frac{H^2\log\frac{K^2 n}{\delta_3}}{\epsilon^2}$ for some constant $C > 0$. The total number of samples used is $KN_0 + N_1 + (K + nK^2)N_2$. Substituting the choices of parameters into N_0, N_1, and N_2, we proved the sample complexity. Furthermore, for time complexity analysis, since our algorithm only calls the `BaseCAMP` and our planning algorithm polynomial number of times, time complexity is also bounded by $(OA)^{C(\gamma^{-4}\log\frac{SHO}{\gamma\alpha})}\log\frac{1}{\delta}$. □

F. Experimental Details

Implementation details on MPE. We train both state-of-the-art centralized-training algorithm MAPPO and decentralized-training algorithm IPPO (Yu et al., 2021) with different information sharing mechanisms by varying the information sharing delay from 0 to ∞. For the centralized MAPPO, we also adopt parameter sharing when agents are homogenous, which is reported as important for improved performance. For decentralized IPPO, we do not enforce any coordination during training among agents. Note that the original algorithm in (Yu et al., 2021) corresponds to the case, where the delay is $d = \infty$.

Implementation details on our approaches. Furthermore, for scalability and compatibility with popular deep RL algorithms, we fit the transition using neural networks instead of the counting methods adopted in Algorithm 7. For the planning oracles used in Algorithm 7, we choose to use Q-learning instead of backward-induction style algorithms as in Algorithm 3, for which we found working very well empirically. Finally, for constructing approximate common information, we used finite memory with a length of 4. Both our algorithm and baselines are trained with 80000 time steps.