Partially Observable Multi-Agent Reinforcement Learning with Information Sharing

Xiangyu Liu[†] Kaiqing Zhang[†]

Abstract

We study provable multi-agent reinforcement learning (RL) in the general framework of partially observable stochastic games (POSGs). To circumvent the known hardness results and the use of computationally intractable oracles, we advocate leveraging the potential *information-sharing* among agents, a common practice in empirical multi-agent RL, and a standard model for multi-agent control systems with communications. We first establish several computational complexity results to justify the necessity of information-sharing, as well as the observability assumption that has enabled quasi-polynomial time and sample single-agent RL with partial observations, for efficiently solving POSGs. Inspired by the inefficiency of planning in the ground-truth model, 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 in quasi-polynomial-time, under the aforementioned assumptions. Furthermore, we develop a partially observable multi-agent RL algorithm whose time and sample complexities are *both* quasi-polynomial. Finally, beyond equilibrium learning, we extend our algorithmic framework to find the *team-optimal solution* in cooperative POSGs, i.e., decentralized partially observable Markov decision processes, a more challenging goal. We establish concrete computational and sample complexities under several common structural assumptions of the model. We hope our study could open up the possibilities of leveraging and even designing different *information structures*, a well-studied notion in control theory, for developing both sample- and computation-efficient partially observable multi-agent RL.

1 Introduction

Recent years have witnessed the 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 the environments and other decision-makers. 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 sometimes referred to

[†]University of Maryland, College Park. Email: {xyliu999,kaiqing}@umd.edu. A preliminary version of the paper has been accepted to the International Conference on Machine Learning (ICML) 2023 (Liu and Zhang, 2023).

as *information asymmetry* in game theory and decentralized decision-making (Behn and Ho, 1968; Milgrom and Roberts, 1987; 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 and Tsitsiklis, 1987; Mundhenk et al., 2000; Jin et al., 2020). The challenge is known to be amplified for multi-agent decentralized decision-making (Witsenhausen, 1968; Tsitsiklis and Athans, 1985). Existing partially observable MARL algorithms with finite-time/sample guarantees either only apply to a small subset of highly structured problems (Zinkevich et al., 2007; Kozuno et al., 2021), or require computationally intractable oracles (Liu et al., 2022b).

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

Can partially observable MARL be made 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 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 detail our contributions below.

Contributions We study provable multi-agent RL under the framework of partially observable stochastic games (POSGs), with potential *information sharing* among agents. First, we establish several computational 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 recent literature, which enabled addressing single-agent partially observable RL without computationally intractable oracles. Second, we propose to further approximate the shared common information to construct an *approximate model*, and characterize the computational 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 POSG) has quasi-polynomial time complexity. Third, based on the planning results, we develop a partially observable multi-agent RL algorithm whose time and sample complexities are *both* quasi-polynomial, which we refer to as being *quasi-efficient* for short (given that *polynomial*-complexity algorithms are generally deemed as being *efficient*). Fourth, beyond *equilibrium* learning, we extend our framework and algorithm to find the *team-optimal* solution in cooperative POSGs, i.e., decentralized partially observable Markov decision processes (Dec-POMDPs), a much more challenging goal. To this end, we identify several common structural assumptions on the model under which quasi-efficient planning and learning become attainable. To the best of our knowledge, this is the first study of provable partially observable MARL with (quasi-)efficiency, with *both* sample and computational complexities. Key to our results is to carefully incorporate insights from both *information structures/sharing*, a well-studied framework in decentralized stochastic control theory, and the tractability conditions investigated in recent (single-agent) reinforcement learning theory.

1.1 Related Work

Information sharing in theory and practice. The idea of information-sharing and the study of more general *information structures* has been extensively studied in decentralized stochastic control

(Witsenhausen, 1971; Nayyar et al., 2010, 2013b), as well as dynamic games (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) provided significant inspiration for our work. The information-sharing structures in these works have enabled *backward-induction-based* planning algorithms even in this decentralized setting. Performance bounds of information compression in such a framework were later derived in Mao et al. (2020); Kao and Subramanian (2022). However, no computation nor sample complexities of algorithms were analyzed in these works. On the other hand, information-sharing has become a normal practice in empirical MARL (Lowe et al., 2017; Sunehag et al., 2018; Rashid et al., 2020), usually instantiated via the so-called *centralized training*, where all agents' information was shared in the training to improve learning efficiency. However, information-sharing/structure has not been fully investigated in the theoretical studies of MARL.

Decentralized stochastic control and decision-making. Decentralized stochastic control and decision-making are known to have unique challenges, compared to the single-agent and centralized counterpart, since the seminal works Witsenhausen (1968); Tsitsiklis and Athans (1985). In particular, Tsitsiklis and Athans (1985) showed that variations of the classical "team decision problem" can be `NP-hard`. Later, Bernstein et al. (2002) showed that planning in Dec-POMDPs, a special class of POSGs with an identical reward function shared across agents, can be `NEXP-hard` in finding the team-optimal solution. Hansen et al. (2004) provided 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 and Zilberstein, 2009; Horák et al., 2017).

RL in partially observable environments. It is known that in general, planning in even *single-agent* POMDPs can be `PSPACE-complete` (Papadimitriou and 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 recent literature on RL in POMDPs with additional assumptions, e.g., Azizzadenesheli et al. (2016); Jin et al. (2020); Liu et al. (2022a). However, these works only focused on *statistical* efficiency, and the algorithms usually required *computationally intractable* oracles. More recently, Golowich et al. (2022b) has identified the condition of γ-observability in POMDPs (firstly introduced in Even-Dar et al. (2007)), which enabled a quasi-polynomial-time planning algorithm for such POMDPs. Subsequently, Golowich et al. (2022a) has developed an RL algorithm based on the planning one in Golowich et al. (2022b), which was both sample and computation (quasi-)efficient. The key enabler of these (quasi-)efficient algorithms is the use of the *finite-memory* policy class, whose (near-)optimality has also been studied lately in Kara (2022); Kara and Yüksel (2022), under different assumptions on both the transition dynamics and the observation channels. Rather than statistical and computational complexity guarantees, Subramanian et al. (2022) has analyzed the performance bounds of general approximate information states (AIS) in partially observable environments. Our finite-memory compression may be viewed as a kind of AIS, although it does not satisfy the *uniform* approximation conditions in Subramanian et al. (2022) (and also in Mao et al. (2020); Kao and Subramanian (2022)). In fact, relaxing such conditions to *expected* versions is the key to obtaining our (quasi-)efficient sample and computational complexities (c.f. Remark 2). Other information compression results include Tang et al. (2024) for dynamic games, and Sinha and Mahajan (2023); Cai et al. (2024) for RL with asymmetric information.

Provable multi-agent reinforcement learning. There has been a fast-growing literature on provable MARL algorithms with sample efficiency guarantees, e.g., Bai et al. (2020); Liu et al. (2021); Zhang et al. (2020); Xie et al. (2020); Zhang et al. (2021b); Wei et al. (2021); Daskalakis et al. (2020); Jin et al. (2024); Song et al. (2021); Daskalakis et al. (2022); Mao et al. (2022); Leonardos et al. (2022);

Zhang et al. (2021c); Ding et al. (2022); Chen et al. (2023). However, these works have been exclusively focused on the fully observable setting of 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 algorithm in Kozuno et al. (2021) only applied to POSGs with certain tree-structured transitions, while that in Liu et al. (2022b) required computationally intractable oracles. In general, information-sharing/structure has not been fully investigated in the theoretical studies of MARL with *finite-sample* and *computation complexities*. One exception is Kao et al. (2022), which exploited a special *hierarchical* information structure in the bandits and MDP settings. Another exception is Altabaa and Yang (2024), which appeared online after the acceptance of the conference version of this paper (Liu and Zhang, 2023), and also incorporated (general) *information structure* considerations into the algorithm design and analyses. However, the algorithms in Altabaa and Yang (2024) also required computationally intractable oracles, with a focus on statistical-tractability only.

Independent result in Golowich et al. (2023). We note that after the acceptance to ICML 2023 of the preliminary version for the paper, an updated version of Golowich et al. (2022b) in its proceedings form appeared online, i.e., Golowich et al. (2023). In Golowich et al. (2023), a quasi-polynomial-time planning algorithm for solving a class of partially observable stochastic games was also discussed. There are several differences compared to our results. First, in the class of POSGs considered in Golowich et al. (2023), the observation is identical for all the players, and each player has access to the joint action history of all the players. Interestingly, this setting exactly corresponds to the *fully-sharing/symmetric-information* case covered by our information-sharing framework (see Example 3 in Appendix B). Moreover, we study both Nash equilibrium (NE) in cooperative/zero-sum games and correlated equilibrium (CE), coarse correlated equilibrium (CCE) in general-sum games, while Golowich et al. (2023) only focused on finding CCE in general-sum games; we also establish a result for *learning* equilibria in POSGs with both quasi-polynomial sample and computational complexities, while Golowich et al. (2023) only focused on planning with model knowledge. Additionally, we also establish results for team-optimum learning for Dec-POMDPs under certain structural conditions.

Notation. For two sets B and D, we define $B \setminus D$ as the set of elements that are in B but not in 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 $x_{\max\{a,m\}:\min\{b,n\}}$. For an event $\mathcal{E}$, we use $\mathbf{1}$ to denote the indicator function such that $\mathbf{1}(\mathcal{E}) = 1$ if the event $\mathcal{E}$ is true and 0 otherwise. For a finite set B, we let $\Delta(B)$ denote the set of distributions over B. For two probability distributions p, q, we define the 2-Rényi divergence as $D_2(p\|q) := \log \mathbb{E}_{x\sim p}\left[\frac{p(x)}{q(x)}\right]$. We also define $p \ll q$ if $q(x) = 0$ implies $p(x) = 0$.

2 Preliminaries

2.1 POSGs and information sharing

Model. Formally, we define a finite-horizon 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 $\mathcal{S}$ denotes the state space with $|\mathcal{S}| = S$, $\mathcal{A}_i$ denotes the action space for the i^{th} agent with $|\mathcal{A}_i| = A_i$, and H denotes the length of each episode. We denote by $a_h := (a_{1,h}, \cdots, a_{n,h})$ the joint action of all the n agents at time step h, and by $\mathcal{A} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_n$ the joint action space with $|\mathcal{A}| = A = \prod_{i=1}^n A_i$. 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 is 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 the n agents, and by $\mathcal{O} := \mathcal{O}_1 \times \cdots \times \mathcal{O}_n$ with $|\mathcal{O}| = O = \prod_{i=1}^n O_i$. We use $\mathbb{O} = \{\mathbb{O}_h\}_{h \in [H+1]}$ to denote the collection of the joint emission matrices, so that $\mathbb{O}_h(\cdot \mid s) \in \Delta(\mathcal{O})$ 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 $\mathbb{O}_h(\cdot \mid s)$. We also denote by $\mathbb{O}_{i,h}(\cdot \mid s) \in \Delta(\mathcal{O}_i)$ the marginalized emission for the i^{th} agent at state s. Finally, $r_i = \{r_{i,h}\}_{h \in [H]}$ is a collection of reward functions, so that $r_{i,h}(s_h, a_h) \in [0,1]$ is the reward of the i^{th} agent given the state s_h and (joint) action a_h taken at step h. This general formulation of POSGs includes several important subclasses. For example, decentralized partially observable Markov decision processes (i.e., 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 $r_1 + r_2 = 1$. Note that we require $r_1 + r_2$ to be 1 instead of 0 to be consistent with our assumption that $r_{i,h} \in [0,1]$ for each $i \in [0,1]$, and this requirement does not lose any optimality as one can always subtract the constant-sum offset to attain a zero-sum structure. Hereafter, we may use the terminology *cooperative POSG* and *Dec-POMDP* interchangeably.

Information sharing, common and private information. Each agent i at step h in the POSG maintains its own information, $\tau_{i,h}$, a collection of (potentially partial) historical observations and actions at step h, namely, $\tau_{i,h} \subseteq \{o_1, a_1, o_2, \cdots, a_{h-1}, o_h\}$, and the collection of such histories at step h is denoted by $\mathcal{T}_{i,h}$. In many practical examples (see some concrete ones in §B), agents may share part of the history with each other, which may introduce more structures in the game that enable both sample and computation efficient learning. The information sharing splits the full history into the *common/shared* and the *private* information for each agent. The *common information* at step h is a subset of the joint history τ_h: $c_h \subseteq \{o_1, a_1, o_2, \cdots, a_{h-1}, o_h\}$, which is available to *all the agents* in the system, and the collection of the common information is denoted as $\mathcal{C}_h$ and we 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 $P_h = \prod_{i=1}^n P_{i,h}$. We allow c_h or $p_{i,h}$ to take the special value of $\emptyset$ when there is no common or private information. In particular, when $\mathcal{C}_h = \{\emptyset\}$, the problem reduces to a 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, for which the goal is usually to find the team-optimal policy.

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 all z_{h+1} at step h.

- *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, and thus the common information at the next step c_{h+1}, depends on the "new" information $\{a_h, o_{h+1}\}$ generated between steps h and $h+1$ and part of the "old" information p_h. The incremental common information can be generated by certain sharing and communication protocols among 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), specifying the source of common information and private information. Based on such evolution rules, we define $\{f_h\}_{h\in[H+1]}$ and $\{g_h\}_{h\in[H+1]}$, where $f_h : \mathcal{A}^{h-1} \times \mathcal{O}^h \to \mathcal{C}_h$ and $g_h : \mathcal{A}^{h-1} \times \mathcal{O}^h \to \mathcal{P}_h$ for $h \in [H+1]$, as the mappings that map the joint history to common information and joint private information, respectively.

2.2 Policies and value functions

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

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

where Ω_h is a space of random seeds shared among agents. 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), which maps the *available information* of the i^{th} agent, i.e., the private information and the common information, together with potentially local random seed $\omega_{i,h} \in \Omega_h$, to the distribution over her actions. We further denote by $\Pi_i = \times_{h\in[H]}\Pi_{i,h}$ 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 a *deterministic* action for agent i, and denote the joint space of such policies as $\widetilde{\Pi}$.

One 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 Γ_h as the function class of joint prescriptions. Intuitively, the *partial* function $\pi_{i,h}(\cdot\,|\,\omega_{i,h}, \cdot, c_h)$ 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., $\pi_i = \{\pi_{i,1}, \cdots, \pi_{i,H}\}$. 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 distributions of drawing each seed $\omega_{i,h}$ for different agents are *independent*. Furthermore, sometimes, we might resort to *deterministic* joint policies with *joint history* as input (which could potentially go beyond Π): $\pi = \{\pi_1, \pi_2, \cdots, \pi_H\}$, where π_h is defined as: $\pi_h : \mathcal{A}^{h-1} \times \mathcal{O}^h \to \mathcal{A}$. We denote the collection of such policies as Π^{det}, and note that $\Pi \subseteq \Delta(\Pi^{\mathrm{det}})$. For any policy $\pi \in \Delta(\Pi^{\mathrm{det}})$ and event $\mathcal{E}$, we write $\mathbb{P}^{\mathcal{G}}_{s_{1:h}, a_{1:h-1}, o_{1:h} \sim \pi_{1:h-1}}(\mathcal{E})$ to denote the probability of $\mathcal{E}$ when $\{s_{1:h}, a_{1:h-1}, o_{1:h}\}$ 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 $\{s_{1:h}, a_{1:h-1}, o_{1:h}\}$ 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. Moreover, we will write $\mathbb{E}^{\mathcal{G}}_{s_{1:h}, a_{1:h-1}, o_{1:h} \sim \pi}[\cdot]$ or $\mathbb{E}^{\mathcal{G}}_{\pi}[\cdot]$ to denote expectations of the trajectories under policy π, 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 $\mathbb{P}_h^{\mathcal{M}}(\cdot)$, $\mathbb{E}_{\pi}^{\mathcal{M}}[\cdot]$, and $\mathbb{E}^{\mathcal{M}}[\cdot]$ are defined in the same way. We will hereafter use *strategy* and *policy* interchangeably. We are now ready to define the *value function* for each agent under our framework:

Definition 1 (Value function with information sharing). *For each agent $i \in [n]$ and step $h \in [H]$, given common information c_h and joint policy $\pi = \{\pi_i\}_{i=1}^n \in \Pi$, the* value function conditioned on the common information *of agent i is defined as:*

$$V_{i,h}^{\pi,\mathcal{G}}(c_h) := \mathbb{E}_{\pi}^{\mathcal{G}}\left[\sum_{h'=h}^{H} r_{i,h'}(s_{h'}, a_{h'}) \,\Big|\, c_h\right],$$

where the expectation is taken over the randomness from the model $\mathcal{G}$, policy π, and the random seeds. For any $c_{H+1} \in \mathcal{C}_{H+1}$: $V_{i,H+1}^{\pi,\mathcal{G}}(c_{H+1}) := 0$. For the value function at the first step, we denote $V_{i,1}^{\pi,\mathcal{G}}(\emptyset) := \mathbb{E}^{\mathcal{G}}[V_{i,h}^{\pi,\mathcal{G}}(c_1)] = \mathbb{E}_{\pi}^{\mathcal{G}}[\sum_{h=1}^{H} r_{i,h}(s_h, a_h)]$, where the expectation is taken over the randomness of c_1, which is a function of o_1 and does not depend on π.

Correspondingly, we can define the prescription-value function $Q_{i,h}^{\pi,\mathcal{G}}(c_h, \gamma_h)$, a generalization of the *action-value* function in MDPs, indicating the expected return for the i^{th} agent when all the agents firstly adopt the prescriptions $\{\gamma_{j,h}\}_{j\in[n]}$ at step h and then follow π (c.f. Definition 12).

2.3 Solution concepts

With the definition of the value functions, we can accordingly define the solution concepts, ϵ-NE (and similarly ϵ-CCE, ϵ-CE), and ϵ-team optimum under the information-sharing framework as follows.

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

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

Definition 4 (ϵ-approximate correlated equilibrium with information sharing)**.** *For any $\epsilon \geq 0$, a joint policy $\pi^\star \in \Pi$ is an ϵ-approximate correlated equilibrium of the POSG $\mathcal{G}$ with information sharing 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 a 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 that this definition extends those in Song et al. (2021); Liu et al. (2021); Jin et al. (2024) to our settings when there exists common information, and is a natural generalization of the definition in the normal-form game case (Roughgarden, 2010).*

Definition 5 (ϵ-approximate team-optimum in Dec-POMDPs with information sharing)**.** *When the reward functions $r_{i,h}$ are identical for all $i \in [n]$, i.e., $r_{i,h} = r_h$, the POSG reduces to a Dec-POMDP, and a policy $\pi^\star \in \widetilde{\Pi}$ is an ϵ-approximate team-optimal policy if: $V_1^{\pi^\star,\mathcal{G}}(\emptyset) \geq \max_{\pi' \in \widetilde{\Pi}} V_1^{\pi',\mathcal{G}}(\emptyset) - \epsilon$, where we have omitted the agent index for the value function.*

It is also worth noting that, under given information-sharing structures, the team-optimal solution is always a NE in the Dec-POMDP setting, and in general, a NE is always a CE, and a CE is always a CCE.

3 Hardness and Planning with Exact Model

3.1 Hardness on finding equilibria

Recently, reference Golowich et al. (2022b) considered *observable* POMDPs (firstly introduced in Even-Dar et al. (2007)) 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 thus make a similar observability assumption on the *joint* emission matrix as below. Note that this is weaker than making the assumption on the *individual* emission matrix of each agent.

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 $h \in [H]$, 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 γ-observability if all its $\mathbb{O}_h$ for $h \in [H]$ do so.

Examples of an observation matrix which satisfies γ-observability include the random channel which outputs the hidden state with probability γ, and otherwise outputs a random state uniformly (i.e., from a "noisy sensor") or an extra dummy observation $\emptyset$ deterministically (i.e., from a "failure mode"). Meanwhile, although the tractability of NE/CE/CCE in normal-form games has been extensively studied, its formal tractability in POSGs has been less studied. Here by the following proposition, we show that both Assumption 2 and some favorable information-sharing structure are *necessary* for NE/CE/CCE to be computationally tractable, even for the special classes of zero-sum POSGs and cooperative POSGs. Specifically, they are necessary in the sense that missing either one of them would make seeking approximate NE/CE/CCE computationally hard, whose proof is deferred to §E.1.

Proposition 1. *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`*.*

Hence, we will now focus on planning and learning under these assumptions.

3.2 Planning with strategy-independent common belief

For both optimal and equilibrium policy computation, it is known that *backward induction* is one of the most useful approaches for solving (fully-observable) stochastic games. However, the essential impediment to applying backward induction in *asymmetric-information/partially observable* dynamic 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 non-zero probability under the trajectories generated by $\pi_{1:h-1}$. Consider any other policies $\pi'_{1:h-1}$, which also give a non-zero probability to c_h. Then, we assume that: for any such $c_h \in \mathcal{C}_h$, and any $p_h \in \mathcal{P}_h, s_h \in \mathcal{S}$, $\mathbb{P}_h^{\pi_{1:h-1},\mathcal{G}}(s_h, p_h \mid c_h) = \mathbb{P}_h^{\pi'_{1:h-1},\mathcal{G}}(s_h, p_h \mid c_h)$.*

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 our case) in Assumption 3 is *independent* of the control strategy. For more detailed discussions, we refer to Nayyar et al. (2013a). Before proceeding with further analysis, we introduced some common examples in §B that satisfy this assumption (see Nayyar et al. (2013a) and also §E.4).

With Assumption 3, we are able to develop the planning algorithm (summarized 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. Note that a value-iteration algorithm for *NE computation* was firstly also studied in Nayyar et al. (2013a), over the space of *common-information-based beliefs* (instead of that of common information). By planning over the common-information space, we can establish its computational complexity, which was not established in Nayyar et al. (2013a), and enable a more efficient planning algorithm later by truncating the common information properly (c.f. §4.1). We now establish the computational complexity of Algorithm 1 more concretely.

Theorem 1. *Fix $\epsilon > 0$. For the POSG $\mathcal{G}$ that satisfies Assumptions 1 and 3, given access to the belief $\mathbb{P}_h^{\mathcal{G}}(s_h, p_h \mid c_h)$, 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 $\max_{h\in[H]} C_h \cdot poly(S, A, P_h, H, \frac{1}{\epsilon})$.*

To prove this, we will prove a more general theorem (see Theorem 2 later), of which Theorem 1 is a special case. This theorem characterizes the dependence of computational complexity on the cardinality of the common information set and private information set. To get a sense of how large C_hP_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 6 (Perfect recall)**.** *We say that player i has perfect recall if for any $h \in [H]$, it holds that $\{a_{i,1:h-1}, o_{i,1:h}\} \subseteq \tau_{i,h}$, and $\tau_{i,h} \subseteq \tau_{i,h+1}$.*

If each player has perfect recall as defined above, we can show that C_hP_h must be exponential in the horizon index h. Proof of the result below can be found in §E.1.

Lemma 1. *Fix any $h \in [H]$, and suppose Assumption 1 holds. Then, if each player has perfect recall as given in Definition 6, then for any information-sharing structure, we have $C_hP_h \geq (OA)^{h-1}$.*

From this result, together with Theorem 1, we know that the computational complexity of such a naive planning algorithm must suffer from the exponential dependence of $\Omega((OA)^h)$. This negative result implies that it is barely possible to get computational efficiency for planning in the true model $\mathcal{G}$, since the cardinality C_hP_h has to be very large oftentimes. Meanwhile, it is worth noting that for obtaining Theorem 1, we have not yet leveraged our Assumption 2. Thus, this negative result is in line with our fundamental hardness results in Proposition 1.

4 Planning and Learning with Approximate Common Information

4.1 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 standard scenario with perfect recall. To address this issue, we propose to further *compress* the information available to the agent under certain regularity conditions, while approximately maintaining the optimality of the policies computed/learned from the compressed information. Notably, there is a trade-off between *compression error* and *computational tractability*. We 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, we first define the *approximate* common information model in our setting.

Definition 7 (Approximate common information model). *We define an* expected approximate common information model *of $\mathcal{G}$ as*

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

*where Γ is the function class for joint prescriptions, $\widehat{\mathcal{C}}_h$ is the space of approximate common information at step h, $\mathbb{P}_h^{\mathcal{M},z} : \widehat{\mathcal{C}}_h \times \Gamma_h \to \Delta(\mathcal{Z}_{h+1})$ gives the probability of z_{h+1} given $\widehat{c}_h \in \widehat{\mathcal{C}}_h$ and $\{\gamma_{i,h}\}_{i\in[n]} \in \Gamma_h$, with $\mathcal{Z}_{h+1}$ being the space of incremental common information. Similarly, $\widehat{r}_{i,h}^{\mathcal{M}} : \widehat{\mathcal{C}}_h \times \Gamma_h \to [0,1]$ gives the reward of agent i at step h 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 $\{\widehat{c}_h\}_{h\in[H+1]}$ for some compression function* Compress_h *that yields $\widehat{c}_h =$* $\text{Compress}_h(c_h)$*, if it satisfies the following:*

- *It evolves in a recursive manner, i.e., for each $h \in [H]$, there exists a transformation function $\widehat{\phi}_{h+1}$ such that*
$$\widehat{c}_{h+1} = \widehat{\phi}_{h+1}(\widehat{c}_h, z_{h+1}), \tag{4.1}$$
where we recall that $z_{h+1} = c_{h+1} \setminus c_h$ is the common information increment.

- *It suffices for approximately evaluating the performance, i.e., for any $i \in [n]$, any prescription $\gamma_h \in \Gamma_h$ and joint policy $\pi' \in \Pi^{\text{det}}$, it holds that*
$$\mathbb{E}^{\mathcal{G}}_{a_{1:h-1},o_{1:h}\sim\pi'}\left|\mathbb{E}^{\mathcal{G}}[r_{i,h}(s_h,a_h)\mid c_h,\gamma_h] - \widehat{r}_{i,h}^{\mathcal{M}}(\widehat{c}_h,\gamma_h)\right| \le \epsilon_r(\mathcal{M}). \tag{4.2}$$

- *It suffices for approximately predicting common information increment: for any $\gamma_h \in \Gamma_h$ $\pi' \in \Pi^{\text{det}}$, and for $\mathbb{P}_h^{\mathcal{G}}(z_{h+1}\,|\,c_h,\gamma_h)$ and $\mathbb{P}_h^{\mathcal{M},z}(z_{h+1}\,|\,\widehat{c}_h,\gamma_h)$, we have*
$$\mathbb{E}^{\mathcal{G}}_{a_{1:h-1},o_{1:h}\sim\pi'}\left\|\mathbb{P}_h^{\mathcal{G}}(\cdot\,|\,c_h,\gamma_h) - \mathbb{P}_h^{\mathcal{M},z}(\cdot\,|\,\widehat{c}_h,\gamma_h)\right\|_1 \le \epsilon_z(\mathcal{M}). \tag{4.3}$$

Remark 1. *The approximate model $\mathcal{M}$ defined above can be treated as a (fully-observable) stochastic game, where the state space is $\{\widehat{\mathcal{C}}_h\}_{h\in[H+1]}$, Γ is the joint action space, the composition of $\{\mathbb{P}_h^{\mathcal{M},z}\}_{h\in[H]}$ and $\{\widehat{\phi}_{h+1}\}_{h\in[H]}$ yields the state transition kernel, and $\widehat{r}_{i,h}^{\mathcal{M}}(\widehat{c}_h,\gamma_h)$ is the reward of agent i given* state *$\widehat{c}_h$ and* joint action *γ_h.*

Remark 2. *Note that related definitions in Kao and Subramanian (2022); Mao et al. (2020); Subramanian et al. (2022) required the* total variation distance *between $\mathbb{P}_h^{\mathcal{G}}(\cdot\,|\,c_h,\gamma_h)$ and $\mathbb{P}_h^{\mathcal{M},z}(\cdot\,|\,\widehat{c}_h,\gamma_h)$ to be* uniformly *bounded for* all *c_h. In fact, this kind of compression may be unnecessary and computationally intractable when it comes to efficient planning. Firstly, some common information c_h 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 $c_h = \{a_{1:h-1}, o_{1:h}\}$, the size of such approximate information with errors uniformly bounded for* all *$\{a_{1:h-1}, o_{1:h}\}$ may not be sub-exponential under Assumption 2, 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.*

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 $\mathcal{M}$, i.e., mainly how to define $(\{\mathbb{P}_h^{\mathcal{M},z}\}_{h\in[H]}, \widehat{r}^{\mathcal{M}})$, even if we are already given certain compression functions. To address this, in the following, we provide a way to construct $(\{\mathbb{P}_h^{\mathcal{M},z}\}_{h\in[H]}, \widehat{r}^{\mathcal{M}})$ from an approximate belief over the states and private information $\{\mathbb{P}_h^{\mathcal{M},c}(s_h, p_h \mid \widehat{c}_h)\}_{h\in[H]}$.

Definition 8 (Model-belief consistency)**.** *We say the expected approximate common information model $\mathcal{M}$ is* consistent *with some belief $\{\mathbb{P}_h^{\mathcal{M},c}(s_h, p_h \mid \widehat{c}_h)\}_{h\in[H]}$ if it satisfies the following for all $i \in [n]$, $h \in [H]$:*

$$\mathbb{P}_h^{\mathcal{M},z}(z_{h+1} \mid \widehat{c}_h, \gamma_h) = \sum_{\substack{s_h, p_h, a_h, o_{h+1}: \\ \chi_{h+1}(p_h, a_h, o_{h+1}) = z_{h+1}}} \left(\mathbb{P}_h^{\mathcal{M},c}(s_h, p_h \mid \widehat{c}_h) \prod_{j=1}^{n} \gamma_{j,h}(a_{j,h} \mid p_{j,h}) \times \sum_{s_{h+1}} \mathbb{T}_h(s_{h+1} \mid s_h, a_h) \mathbb{O}_{h+1}(o_{h+1} \mid s_{h+1})\right), \tag{4.4}$$

$$\widehat{r}_{i,h}^{\mathcal{M}}(\widehat{c}_h, \gamma_h) = \sum_{s_h, p_h, a_h} \mathbb{P}_h^{\mathcal{M},c}(s_h, p_h \mid \widehat{c}_h) \prod_{j=1}^{n} \gamma_{j,h}(a_{j,h} \mid p_{j,h}) r_{i,h}(s_h, a_h). \tag{4.5}$$

With such an expected approximate common information model, similar to Algorithm 1, we develop a value-iteration-type algorithm (see pseudocode in Algorithm 3) running on the model $\mathcal{M}$ instead of $\mathcal{G}$, which outputs an approximate NE/CE/CCE, enjoying the following guarantees. The key benefit of requiring the model $\mathcal{M}$ to be *consistent* with some belief is that under this condition, the stage game in Algorithm 3 can be formulated as a *multi-linear* game of *polynomial* size, thus computing its equilibrium is computationally tractable (c.f. §E.2).

Theorem 2. *Fix $\epsilon_r, \epsilon_z, \epsilon_e > 0$. Given any (ϵ_r, ϵ_z)-expected-approximate common information model $\mathcal{M}$ for the POSG $\mathcal{G}$ under Assumptions 1 and 3. Furthermore, if $\mathcal{M}$ is consistent with some given approximate belief $\{\mathbb{P}_h^{\mathcal{M},c}(s_h, p_h \mid \widehat{c}_h)\}_{h\in[H]}$ (in the sense of Definition 8), then there exists an algorithm, Algorithm 3, that can output an ϵ-NE if $\mathcal{G}$ is zero-sum or cooperative, or ϵ-CE/CCE if $\mathcal{G}$ is general-sum, where $\epsilon := 2H\epsilon_r + H^2\epsilon_z + H\epsilon_e$ with time complexity $\max_{h\in[H]} \widehat{C}_h \cdot poly(S, A, P_h, H, \frac{1}{\epsilon_e})$.*

As a sanity check, by choosing the compression function as the identity mapping, Theorem 2 recovers Theorem 1.

Planning in observable POSGs without intractable oracles. Theorem 2 applies to any expected approximate common information model as given in Definition 7, 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 computationally 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, for all the information-sharing structures in §B, there exists a quasi-polynomial time algorithm that can compute an ϵ-NE if $\mathcal{G}$ is zero-sum or cooperative, and an ϵ-CE/CCE if $\mathcal{G}$ is general-sum.*

4.2 Statistically (quasi-)efficient learning

Until now, we have been assuming the full knowledge of the model $\mathcal{G}$ (the transition kernel, emission, and reward functions). In this full-information setting, we are able to construct some model $\mathcal{M}$ to approximate the true model $\mathcal{G}$ according to the conditions we identified in Definition 7. However, when we only have access to the samples drawn from the POSG $\mathcal{G}$, it is difficult to directly construct such a model $\mathcal{M}$ due to the lack of the model specification. To address this issue, we propose to construct a specific expected approximate common information model that *depends on the policies* $\pi^{1:H}$ *that generate the data for such a construction, which is 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 $\mathcal{G}$. The choice of $\pi^{1:H}$ will be specified later to ensure $\widetilde{\mathcal{M}}(\pi^{1:H})$ to be a good approximation of $\mathcal{G}$.

Compared to Golowich et al. (2022a), there are several key technical challenges our analysis addresses: firstly, Golowich et al. (2022a) only considered *finite-memory* approximation for POMDPs, where the sample complexity can be easily characterized by the *length of the finite memory*. In contrast, our goal is to deal with a more general compression scheme, for which we need to define a generalized quantity that can characterize the sample complexity under general compression schemes (c.f. Definition 10). Secondly and more importantly, Golowich et al. (2022a) essentially learned the transition and reward of the approximate model by simply *enumerating* all possible actions, which corresponds to enumerating all possible prescriptions $\gamma_h \in \Gamma_h$ for each $\widehat{c}_h \in \widehat{\mathcal{C}}_h$ to learn $\mathbb{P}_h^{\widetilde{\mathcal{M}}(\pi^{1:H}),z}(\cdot\,|\,\widehat{c}_h, \gamma_h)$ and $\widehat{r}_{i,h}^{\widetilde{\mathcal{M}}(\pi^{1:H})}(\widehat{c}_h, \gamma_h)$, if one naively applies its algorithm and analyses to our setting. This will lead to an *exponential* sample complexity since even the number of possible *deterministic* prescriptions is A^{P_h} (while all possible *randomized* prescriptions are even larger and infinitely many). To address this challenge, we identify a *decomposition* on the aforementioned quantities to separately learn the distributions of private information and the next observation. Analyzing such a separate learning procedure requires a careful examination of those $\widehat{c}_h \in \widehat{\mathcal{C}}_h$ and $p_h \in \mathcal{P}_h$ that are rarely visited by π^h.

To introduce the aforementioned approximate model $\widetilde{\mathcal{M}}(\pi^{1:H})$, we present the following definition, where the key is to introduce an approximate common information-based belief $\{\mathbb{P}_h^{\pi^h,\mathcal{G}}(s_h, p_h\,|\,\widehat{c}_h)\}_{h\in[H]}$, which is generated by running a certain policy $\pi^h \in \Delta(\Pi^{\mathrm{det}})$ under the true model $\mathcal{G}$.

Definition 9 (Policy-dependent approximate common information model)**.** *Given a model* $\widetilde{\mathcal{M}}$ *(as in Definition 7) and* H *joint policies* $\pi^{1:H}$*, where each* $\pi^h \in \Delta(\Pi^{\mathrm{det}})$ *for* $h \in [H]$*, we say* $\widetilde{\mathcal{M}}$ *is a* policy-dependent expected approximate common information model, *denoted as* $\widetilde{\mathcal{M}}(\pi^{1:H})$*, if 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 8).*

Now we present the main theorem for learning under an expected approximate common information model $\widetilde{\mathcal{M}}(\pi^{1:H})$. A major difference from the analysis for planning in §4.1 is that, we need to *explore* the space of approximate common information, which is a function of a sequence of observations and actions, and we propose to characterize the *length* of the approximate common information as defined below.

Definition 10 (Length of approximate common information)**.** *Given the compression functions* $\{Compress_h\}_{h\in[H+1]}$*, we define the integer* $\widehat{L} > 0$ *as the minimum length such that there exists a mapping* $\widehat{f}_h : \mathcal{A}^{\min\{\widehat{L},h\}} \times \mathcal{O}^{\min\{\widehat{L},h\}} \to \widehat{\mathcal{C}}_h$ *such that for each* $h \in [H+1]$ *and joint history* $\{o_{1:h}, a_{1:h-1}\}$*, we have* $\widehat{f}_h(x_h) = \widehat{c}_h$*, where* $x_h = \{a_{\max\{h-\widehat{L},1\}}, o_{\max\{h-\widehat{L},1\}+1}, \cdots, a_{h-1}, o_h\}$*.*

Such an $\widehat{L}$ will help characterize our final sample complexity, since we need to do exploration for the steps after $h-\widehat{L}$, and $\widehat{L}$ characterizes the cardinality of the space to be explored. With this definition of $\widehat{L}$, we develop Algorithm 6, which learns the model $\widetilde{\mathcal{M}}(\pi^{1:H})$, mainly learning the two quantities $\mathbb{P}^{\widetilde{\mathcal{M}}(\pi^{1:H}),z}$ and $\widehat{r}^{\widetilde{\mathcal{M}}(\pi^{1:H})}$, by executing policies $\pi^{1:H}$ in the true model $\mathcal{G}$, with the following sample complexity.

Theorem 4. *Suppose the POSG $\mathcal{G}$ satisfies Assumptions 1 and 3. Given H policies $\pi^{1:H}$, $\widetilde{\mathcal{M}}(\pi^{1:H})$, and $\widehat{L}$ as in Definition 10, where each $\pi^h \in \Delta(\Pi^{\mathrm{det}})$, $\pi^h_{h-\widehat{L}:h} = \mathrm{Unif}(\mathcal{A})$ for $h \in [H]$. Fix the parameters $\delta_1, \theta_1, \theta_2, \zeta_1, \zeta_2, \epsilon_e > 0$ for Algorithm 6, and some $\phi > 0$, define the approximation error for estimating $\widetilde{\mathcal{M}}(\pi^{1:H})$ using samples under the policies $\pi^{1:H}$ as $\epsilon_{apx}(\pi^{1:H}, \widehat{L}, \zeta_1, \zeta_2, \theta_1, \theta_2, \phi)$. Then, Algorithm 6, can learn 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 $1-\delta_1$, with a sample complexity $N_0 = \mathit{poly}(\max_{h\in[H]} P_h, \max_{h\in[H]} \widehat{C}_h, H, A, O, \frac{1}{\zeta_1}, \frac{1}{\zeta_2}, \frac{1}{\theta_1}, \frac{1}{\theta_2}) \cdot \log\frac{1}{\delta_1}$, where $\epsilon := H\epsilon_r(\widetilde{\mathcal{M}}(\pi^{1:H})) + H^2\epsilon_z(\widetilde{\mathcal{M}}(\pi^{1:H})) + (H^2+H)\epsilon_{apx}(\pi^{1:H}, \widehat{L}, \zeta_1, \zeta_2, \theta_1, \theta_2, \phi) + H\epsilon_e$.*

A detailed version of the theorem is in §D.2. This meta-theorem establishes a sample complexity guarantee of learning expected approximate common information model $\widetilde{\mathcal{M}}(\pi^{1:H})$ in a online exploration setting, which holds for *any* compression functions and policies $\pi^{1:H}$, whose choices are specified next.

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

Theorem 5. *Under Assumption 2, for all the information-sharing structures in §B, 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 probability at least $1-\delta$, with both quasi-polynomial time and sample complexities $(AO)^{C\gamma^{-4}\log\frac{SHO}{\gamma\epsilon}}\log\frac{1}{\delta}$ for some universal constant $C > 0$.*[1]

Due to space constraints, a detailed version of the theorem is presented in §D.2, 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 relied on computationally intractable oracles.

5 Finding Team-Optimum in Dec-POMDPs

Until now, we have primarily focused on solving *equilibria* in POSGs. One notable subclass of POSGs are the Dec-POMDPs, for which a stronger (than equilibrium) solution concept of *team-optimum* (c.f. Definition 5) is usually preferred. Our algorithmic framework developed in §4.1 for planning can be readily extended to computing the team optimal solution, where the only modification is to replace the *equilibrium-computation* subroutine at each step h over the prescription space in Algorithm 3 by a *joint-maximization* one over the prescription space. Specifically, we only need to replace the line 9 of Algorithm 3 by its line 11, i.e., the following step:

$$\{\pi_{1,h}(\cdot\,|\,\widehat{c}_h,\cdot),\cdots,\pi_{n,h}(\cdot\,|\,\widehat{c}_h,\cdot)\} \leftarrow \arg\max_{\gamma_{1,h},\cdots,\gamma_{n,h}} Q_h^{\star,\mathcal{M}}(\widehat{c}_h,\gamma_{1,h},\cdots,\gamma_{n,h}), \tag{5.1}$$

where we omit the agent index for the Q-function, since all the agents share the same Q-function for the Dec-POMDP setting.

Unfortunately, although we can show in Proposition 8 that such a Q-value is *linear* w.r.t. each $\gamma_{i,h}$, it is not necessarily *concave* w.r.t. $\{\gamma_{1,h},\cdots,\gamma_{n,h}\}$ *jointly*. Thus, implementing this maximization subroutine can be computationally intractable. In fact, it is an `NP-hard` problem without additional assumptions.

[1]Note that throughout the paper, we regard the delay d in the examples in Appendix B as a *constant*. In fact, as shown in the full version of the theorem in §D.2, d is allowed to grow *logarithmically* with the horizon H without changing the order of the computational or sample complexities.

Proposition 2. *Without additional assumptions, even with $n = 2$ agents, solving Equation (5.1) is* `NP-hard`*.*

Proof of Proposition 2 is deferred to §E.7. Hence, it seems hopeless to solve Equation (5.1) efficiently. Fortunately, many Dec-POMDPs in real-world applications enjoy certain structures that can be exploited for efficient computation. Specifically, we identify several (sets of) assumptions below, under which solving Equation (5.1) can be computationally tractable. Note that since we need to do planning in the approximate model $\mathcal{M}$, which is oftentimes constructed based on the *original model* $\mathcal{G}$ and some *approximate belief* $\{\mathbb{P}_h^{\mathcal{M},c}(s_h, p_h \mid \widehat{c}_h)\}_{h\in[H]}$, we will necessarily need assumptions on these two quantities, for which we refer to as the **Part (1)** and **Part (2)** of the assumptions below, respectively.

Condition 1: Turn-based structures **Part (1).** For $\mathcal{G}$, we assume that at each step h, there is only one agent, denoted as, $\mathrm{ctt}(h) \in [n]$ that can affect the state transition. Hence, the transition dynamics take the forms of $\mathbb{T}_h : \mathcal{S} \times \mathcal{A}_{\mathrm{ctt}(h)} \to \mathcal{S}$. Meanwhile, since only agent $\mathrm{ctt}(h)$ can affect the transition, we assume the increment of the common information z_{h+1} is only a function of $(p_h, a_{\mathrm{ctt}(h),h}, o_{h+1})$, i.e., $z_{h+1} = \chi_{h+1}(p_h, a_{\mathrm{ctt}(h),h}, o_{h+1})$ instead of $\chi_{h+1}(p_h, a_h, o_{h+1})$. In other words, since at step h, agents other than $\mathrm{ctt}(h)$ do not affect the transition, we assume their actions are not shared. For the reward, we additionally assume that the reward function has an additive structure, i.e., $r_h(s_h, a_h) = \sum_{j\in[n]} r_{j,h}(s_h, a_{j,h})$ for some functions $\{r_{j,h}\}_{j\in[n]}$. **Part (2).** For the approximate belief, we do not impose any assumption. Note that such turn-based structures have been common in the (fully-observable) stochastic game settings (Filar and Vrieze, 2012; Bai and Jin, 2020).

Condition 2: Nested information-sharing. **Part (1).** For $\mathcal{G}$, we do not impose any assumption. **Part (2).** For the approximate belief, we assume that all the agents form a *hierarchy* according to the private information they possess. Without loss of generality, we assume for each $i, j \in [n]$ such that for $j < i$, it holds that $p_{j,h} = Y_h^{ij}(p_{i,h})$ for some deterministic function Y_h^{ij}. More formally, the approximate belief satisfies that $\mathbb{P}_h^{\mathcal{M},c}(p_{j,h} = Y_h^{ij}(p_{i,h}) \mid p_{i,h}, \widehat{c}_h) = 1$, where $\mathbb{P}_h^{\mathcal{M},c}(p_{j,h} \mid p_{i,h}, \widehat{c}_h)$ is the *posterior* distribution induced by the joint distribution $\mathbb{P}_h^{\mathcal{M},c}(s_h, p_h \mid \widehat{c}_h)$. In other words, the σ-algebra generated by the private information of agent i includes that of agent j. This structure has also been studied in Peralez et al. (2024) with a heuristic search approach.

Condition 3: Factorized structures. **Part (1).** For $\mathcal{G}$, we assume that the state s_h at each step $h \in [H]$ can be partitioned into n local states, i.e., $s_h = (s_{1,h}, s_{2,h}, \cdots, s_{n,h})$. Meanwhile, the transition kernel takes the product form of $\mathbb{T}_h(s_{h+1} \mid s_h, a_h) = \prod_{i=1}^n \mathbb{T}_{i,h}(s_{i,h+1} \mid s_{i,h}, a_{i,h})$, the emission also takes the product form of $\mathbb{O}_h(o_h \mid s_h) = \prod_{i=1}^n \mathbb{O}_{i,h}(o_{i,h} \mid s_{i,h})$, and the reward function can be decoupled into n terms such that $r_h(s_h, a_h) = \sum_{i,h} r_{i,h}(s_{i,h}, a_{i,h})$. **Part (2).** For the approximate belief, we assume the constructed approximate common information is also factorized so that $\widehat{c}_h = (\widehat{c}_{1,h}, \cdots, \widehat{c}_{n,h})$, and its evolution additionally satisfies that $\widehat{c}_{i,h+1} = \widehat{\phi}_{i,h+1}(\widehat{c}_{i,h}, z_{i,h+1})$ for some function $\widehat{\phi}_{i,h+1}$ instead of the original Equation (4.1). Correspondingly, the approximate belief needs to satisfy that $\mathbb{P}_h^{\mathcal{M},c}(s_h, p_h \mid \widehat{c}_h) = \prod_{i=1}^n \mathbb{P}_{i,h}^{\mathcal{M},c}(s_{i,h}, p_{i,h} \mid \widehat{c}_{i,h})$, for some functions $\{\mathbb{P}_{i,h}^{\mathcal{M},c}\}_{i\in[n],h\in[H]}$.

Under each of these conditions, Equation (5.1) can be solved exactly with time complexity $\mathtt{poly}(S, A, P_h)$. The key insight into why these conditions suffice is that, they make solving the *joint maximization* in Equation (5.1) equivalent to either solving *individual maximization* for each agent, or *sequential maximization* across agents that can be solved via dynamic programming. Formal statements can be found in Proposition 11, Proposition 12, and Proposition 13 for each condition, respectively. The key insight into why these conditions suffice is that, they make solving the *joint maximization* in Equation (5.1) equivalent to either solving *individual maximization* for each agent, or *sequential maximization* across agents that can be solved via dynamic programming. Once Equation (5.1) can be solved computationally efficiently, computation of the team optimum of Dec-POMDPs becomes tractable, under the same algorithmic framework as in §4.1.

Theorem 6. *Fix $\epsilon > 0$, and consider a Dec-POMDP $\mathcal{G}$ satisfying Assumption 2, then all the examples in §B except the one-step delayed sharing case satisfy either **Condition 1** or **Condition 2** in §5. Hence, Equation (5.1) can be solved in time complexity $\texttt{poly}(S, A, P_h)$ for these cases. Correspondingly, there exists a quasi-polynomial time algorithm that can compute an ϵ-team optimal policy of $\mathcal{G}$. For the one-step delayed sharing case, if one additionally assumes that the Dec-POMDP $\mathcal{G}$ satisfies* **Part (1)** *of **Condition 3**, then there also exists a quasi-polynomial time algorithm that can compute an ϵ-team optimal policy of $\mathcal{G}$, and moreover, the time complexity is polynomial (instead of exponential) in the number of agents n.*

Extensions to learning settings without model knowledge. With the planning oracle for Dec-POMDPs developed above, our framework of *learning* in POSGs can be readily extended to learning in Dec-POMDPs accordingly, achieving *both* quasi-polynomial time and sample complexities for learning the approximate *team-optimal* policy. Due to space constraints, we defer the detailed results to §E.7.

6 Proof Outlines

In this section, we present the proof outlines of the main results introduced before. More details can be found in the Appendices.

6.1 Proof Outline of Theorem 2

For notational simplicity, we present our proof outline for the NE/CCE case, and the CE case can be derived similarly.

Step 1: Bounding the value difference between $\mathcal{G}$ and $\mathcal{M}$. Since what we care about in the end is the equilibrium gap in the actual game $\mathcal{G}$, we bound the value difference between $\mathcal{M}$ and $\mathcal{G}$ in terms of $\epsilon_z(\mathcal{M})$ and $\epsilon_r(\mathcal{M})$ first.

Lemma 2. *For any given policy $\pi' \in \Delta(\Pi^{\mathrm{det}})$, $\pi \in \Pi$, and $h \in [H+1]$, we have*

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

Note that this lemma holds for any $\pi \in \Pi$, thus also $\widehat{\pi}^\star$ and its unilaterally deviated policy $\pi_i \times \widehat{\pi}^\star_{-i}$, facilitating the following steps.

Step 2: Evaluating the equilibrium gap of $\widehat{\pi}^\star$ under $\mathcal{G}$. Now we are ready to evaluate $\widehat{\pi}^\star$ (the output of Algorithm 3) in $\mathcal{G}$. We define for each agent $i \in [n]$ the *best response* as $\pi_i^\star \in \arg\max_{\pi_i \in \Pi_i} V_{i,1}^{\pi_i \times \widehat{\pi}^\star_{-i},\mathcal{G}}(\emptyset)$. Now for any $\pi' \in \Delta(\Pi^{\mathrm{det}})$:

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

where the second step is from a standard guarantee of value iteration on a (fully-observable) stochastic game $\mathcal{M}$ (Lemma 6) and the third step is by Lemma 2. Letting $h = 1$, we can conclude that NE/CCE-gap$(\widehat{\pi}^\star) \leq 2H\epsilon_r + H^2\epsilon_z + H\epsilon_e$.

Step 3: Analyzing computational complexity. Note that Algorithm 3 is of double-loop type. **(1) For the outer-loop:** it enumerates all $\widehat{c}_h \in \widehat{\mathcal{C}}_h$ at each $h \in [H]$. **(2) For the inner-loop:** the main computation comes from computing the ϵ_e-NE/CE/CCE of the game defined by $\{Q^{\star,\mathcal{M}}_{i,h}(\widehat{c}_h, \cdots)\}_{i\in[n]}$. Note that if we treat this game as a normal-form game with the action being all the deterministic prescriptions, then normal-form game solvers can be plugged in. However, the corresponding time complexity will suffer from the size of the action space, i.e., $A_h^{P_h}$. Instead, we show that if we regard each $\gamma_{i,h}$ as a *concatenation* of simplexes, i.e., $\gamma_{i,h} \in \Delta(\mathcal{A}_i)^{P_{i,h}}$, then $Q^{\star,\mathcal{M}}_{i,h}$ is linear w.r.t. to each individual prescription *under our model-belief consistency condition* (c.f. Definition 8). Thus, an ϵ_e-NE/CE/CCE can be solved with time complexity depending only polynomially on the dimension of $\gamma_{i,h}$, which is $A_i P_{i,h}$, in contrast to the previous $A_h^{P_h}$. By putting the time complexity for the outer-loop and inner-loop together, we obtain the final time complexity.

6.2 Proof Outline of Theorem 3

We take the one-step delayed sharing case as an example, and defer the proofs for other information structures to §E.4.

Step 1: Bounding $\epsilon_r(\mathcal{M})$,$\epsilon_z(\mathcal{M})$ with the belief error. As in Definition 8, one can construct $\mathcal{M}$ from some given compression functions and approximate beliefs of $\{\mathbb{P}^{\mathcal{M},c}_h(s_h, p_h \mid \widehat{c}_h)\}_{h\in[H]}$. Thus, we can relate the model errors of $\mathcal{M}$, i.e., $\epsilon_r(\mathcal{M})$ and $\epsilon_z(\mathcal{M})$, with the error of the approximate belief. Specifically, we show the following.

Lemma 3. *Given any belief $\{\mathbb{P}^{\mathcal{M},c}_h(s_h, p_h \mid \widehat{c}_h)\}_{h\in[H]}$ and an associated consistent (in the sense of Definition 8) expected approximate common information model $\mathcal{M}$, it holds that for any $h \in [H], c_h \in \mathcal{C}_h, \gamma_h \in \Gamma_h$:*

$$\left\|\mathbb{P}^{\mathcal{G}}_h(\cdot \mid c_h, \gamma_h) - \mathbb{P}^{\mathcal{M},z}_h(\cdot \mid \widehat{c}_h, \gamma_h)\right\|_1 \leq \left\|\mathbb{P}^{\mathcal{G}}_h(\cdot,\cdot \mid c_h) - \mathbb{P}^{\mathcal{M},c}_h(\cdot,\cdot \mid \widehat{c}_h)\right\|_1, \tag{6.1}$$

$$\left|\mathbb{E}^{\mathcal{G}}[r_{i,h}(s_h, a_h) \mid c_h, \gamma_h] - \widehat{r}^{\mathcal{M}}_{i,h}(\widehat{c}_h, \gamma_h)\right| \leq \left\|\mathbb{P}^{\mathcal{G}}_h(\cdot,\cdot \mid c_h) - \mathbb{P}^{\mathcal{M},c}_h(\cdot,\cdot \mid \widehat{c}_h)\right\|_1, \tag{6.2}$$

where we recall that $\widehat{c}_h := \text{Compress}_h(c_h)$, with $\{\text{Compress}_h\}_{h\in[H+1]}$ from $\mathcal{M}$.

Step 2: Compressing common information using finite-memory truncation. Now it remains to design the compression functions $\{\text{Compress}_h\}_{h\in[H+1]}$ and define the associated approximate beliefs $\mathbb{P}^{\mathcal{M},c}_h : \widehat{\mathcal{C}}_h \to \Delta(\mathcal{S} \times \mathcal{P}_h)$ for $h \in [H]$. Specifically, the information structure satisfies $c_h = \{a_{1:h-1}, o_{1:h-1}\}$, $p_{i,h} = \{o_{i,h}\}$, $z_{h+1} = \{o_h, a_h\}$. More importantly, the ground-truth belief can be computed as $\mathbb{P}^{\mathcal{G}}_h(s_h, p_h \mid c_h) = \boldsymbol{b}_h(a_{1:h-1}, o_{1:h-1})(s_h)\mathbb{O}_h(o_h \mid s_h)$, where $\boldsymbol{b}_h$ denotes the posterior state distribution given $(a_{1:h-1}, o_{1:h-1})$ (c.f. formal definition in Definition 11). Fix an integer $L > 0$, we construct the compression of c_h as $\widehat{c}_h = \{a_{h-L:h-1}, o_{h-L+1:h-1}\}$. The approximate belief can be defined similarly as above as

$$\mathbb{P}^{\mathcal{M},c}_h(s_h, p_h \mid \widehat{c}_h) = \boldsymbol{b}'_h(a_{h-L:h-1}, o_{h-L+1:h-1})(s_h)\mathbb{O}_h(o_h \mid s_h),$$

where $\boldsymbol{b}'_h$ denotes the approximate belief state, where one ignores the history before step $h - L$ and performs the belief update via Bayes rule along the trajectory after it from a prior distribution of *uniform distribution on the state* (c.f. detailed definition in Definition 11). Now we are ready to verify that Definition 7 is satisfied.

- Obviously, it satisfies condition (4.1).
- For any $c_h \in \mathcal{C}_h$ and the corresponding $\widehat{c}_h$ constructed above:

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

  By setting $L \geq C\gamma^{-4}\log(\frac{S}{\epsilon})$, according to Equation (E.8) of Theorem 10, we conclude that for any $\pi' \in \Pi^{\mathrm{det}}, h \in [H]$:

$$\mathbb{E}_{\pi'}^{\mathcal{G}}\left\|\boldsymbol{b}_h(a_{1:h-1},o_{1:h-1})-\boldsymbol{b}'_h(a_{h-L:h-1},o_{h-L+1:h-1})\right\|_1 \leq \epsilon.$$

  Note that this result is adapted from the belief contraction result under the γ-observability of Golowich et al. (2022b) for single-agent settings. Therefore, conditions (4.2), (4.3) in Definition 7 are satisfied using Lemma 3 with $\epsilon_r = \epsilon_z = \epsilon$.

Finally, together with Theorem 2, by choosing $L = \mathcal{O}\Big(\gamma^{-4}\log(\frac{SH}{\epsilon})\Big)$, $\epsilon_e = \mathcal{O}(\epsilon/H)$, we proved that $\widehat{\pi}^\star$ is an ϵ-NE/CCE. Meanwhile, it is direct to see that $\widehat{C}_h \leq (AO)^L$ and $P_h \leq O$, thus proving the overall quasi-polynomial time complexity via Theorem 2.

6.3 Proof Outline of Theorem 4

Step 1: Decomposing transitions of $\widetilde{\mathcal{M}}(\pi^{1:H})$. Learning $\mathbb{P}_h^{\widetilde{\mathcal{M}}(\pi^{1:H}),z}(z_{h+1}\,|\,\widehat{c}_h,\gamma_h)$ for the model $\widetilde{\mathcal{M}}(\pi^{1:H})$ is equivalent to learning $\mathbb{P}_h^{\pi^h_{1:h-1},\mathcal{G}}(z_{h+1}\,|\,\widehat{c}_h,\gamma_h)$, given the definition of $\widetilde{\mathcal{M}}(\pi^{1:H})$ in Definition 9 (c.f. also Proposition 10). As highlighted before, learning $\mathbb{P}_h^{\widetilde{\mathcal{M}}(\pi^{1:H}),z}(z_{h+1}\,|\,\widehat{c}_h,\gamma_h)$ by enumerating all $\widehat{c}_h$ and γ_h is *not* statistically efficient if naively following that of Golowich et al. (2022a). To circumvent this issue, we notice

$$\mathbb{P}_h^{\pi^h_{1:h-1},\mathcal{G}}(z_{h+1}\,|\,\widehat{c}_h,\gamma_h)=\sum_{\substack{p_h,a_h,o_{h+1}:\\ \chi_{h+1}(p_h,a_h,o_{h+1})=z_{h+1}}}\mathbb{P}_h^{\pi^h_{1:h-1},\mathcal{G}}(p_h,a_h,o_{h+1}\,|\,\widehat{c}_h,\gamma_h),$$

where we recall χ_{h+1} from Assumption 1. Now we notice the decomposition:

$$\mathbb{P}_h^{\pi^h_{1:h-1},\mathcal{G}}(p_h,a_h,o_{h+1}\,|\,\widehat{c}_h,\gamma_h)=\mathbb{P}_h^{\pi^h_{1:h-1},\mathcal{G}}(p_h\,|\,\widehat{c}_h)\gamma_h(a_h\,|\,p_h)\mathbb{P}_h^{\pi^h_{1:h-1},\mathcal{G}}(o_{h+1}\,|\,\widehat{c}_h,p_h,a_h),$$

where we use the shorthand notation $\gamma_h(a_h\,|\,p_h) := \prod_{i=1}^n \gamma_{i,h}(a_{i,h}\,|\,p_{i,h})$. With such a decomposition, it suffices to learn $\mathbb{P}_h^{\pi^h_{1:h-1},\mathcal{G}}(p_h\,|\,\widehat{c}_h)$ and $\mathbb{P}_h^{\pi^h_{1:h-1},\mathcal{G}}(o_{h+1}\,|\,\widehat{c}_h,p_h,a_h)$.

Step 2: Bounding the statistical error for learning $\mathbb{P}_h^{\pi^h_{1:h-1},\mathcal{G}}(p_h\,|\,\widehat{c}_h)$ and $\mathbb{P}_h^{\pi^h_{1:h-1},\mathcal{G}}(o_{h+1}\,|\,\widehat{c}_h,p_h,a_h)$. The accuracy and sample complexity of learning those two conditional probabilities depend on the visitation probability of $\widehat{c}_h$ and p_h under π^h. Therefore, we firstly handle those $\widehat{c}_h$ and p_h with large visitation probability as follows.

Lemma 4. *Fix $\delta_1, \zeta_1, \zeta_2, \theta_1, \theta_2 > 0$. Given the compression functions and correspondingly the $\widehat{L}$ as per Definition 10, for Algorithm 5, suppose for all $h \in [H]$, $\pi^h \in \Delta(\Pi^{\mathrm{det}})$ satisfies that $\pi^h_{h-\widehat{L}:h} = Unif(\mathcal{A})$, then as long as N_0 in Algorithm 5 satisfies $N_0 \geq poly(A, O, \max_h P_h, \max_h \widehat{C}_h, \frac{1}{\zeta_1}, \frac{1}{\zeta_2}, \frac{1}{\theta_1}, \frac{1}{\theta_2}, \log\frac{1}{\delta_1})$, with probability at least $1-\delta_1$, the following holds for each $h \in [H]$:*

- *For $\widehat{c}_h \in \widehat{\mathcal{C}}_h$ such that $\mathbb{P}_h^{\pi^h_{1:h-1},\mathcal{G}}(\widehat{c}_h) \geq \zeta_1$, Algorithm 5 can learn a $\mathbb{P}_h^{\widehat{\mathcal{M}}(\pi^{1:H})}(\cdot\,|\,\widehat{c}_h) \in \Delta(\mathcal{P}_h)$ such that $\sum_{p_h}\left|\mathbb{P}_h^{\widehat{\mathcal{M}}(\pi^{1:H})}(p_h\,|\,\widehat{c}_h) - \mathbb{P}_h^{\pi^h_{1:h-1},\mathcal{G}}(p_h\,|\,\widehat{c}_h)\right| \leq \theta_1$.*
- *For $(\widehat{c}_h, p_h, a_h) \in \widehat{\mathcal{C}}_h \times \mathcal{P}_h \times \mathcal{A}$ such that $\mathbb{P}_h^{\pi^h_{1:h-1},\mathcal{G}}(\widehat{c}_h, p_h) \geq \zeta_2$, Algorithm 5 can learn a $\mathbb{P}_h^{\widehat{\mathcal{M}}(\pi^{1:H})}(\cdot\,|\,\widehat{c}_h, p_h, a_h) \in \Delta(\mathcal{O})$ such that $\sum_{o_{h+1}}\left|\mathbb{P}_h^{\widehat{\mathcal{M}}(\pi^{1:H})}(o_{h+1}\,|\,\widehat{c}_h, p_h, a_h) - \mathbb{P}_h^{\pi^h_{1:h-1},\mathcal{G}}(o_{h+1}\,|\,\widehat{c}_h, p_h, a_h)\right| \leq \theta_2$.*

We refer to the joint of the two bullets above as event $\mathcal{E}_1$.

Step 3: Relating $\widehat{\mathcal{M}}(\pi^{1:H})$ and $\mathcal{G}$ via $\widetilde{\mathcal{M}}(\pi^{1:H})$. To begin with, we evaluate the approximation errors between $\widetilde{\mathcal{M}}(\pi^{1:H})$ and $\widehat{\mathcal{M}}(\pi^{1:H})$ as follows.

Lemma 5. *Given policies $\pi^{1:H}$ such that π^h satisfies the same condition as in Lemma 4, under the event $\mathcal{E}_1$ in Lemma 4, then for any $h \in [H]$, policy $\pi \in \Delta(\Pi^{\mathrm{det}})$, and prescription $\gamma_h \in \Gamma_h$, it holds that for any $\widehat{c}_h \in \widehat{\mathcal{C}}_h$ with $\mathbb{P}_h^{\pi^h_{1:h-1}}(\widehat{c}_h) \geq \zeta_1$*

$$\left|\mathbb{P}_h^{\widetilde{\mathcal{M}}(\pi^{1:H}),z}(z_{h+1}\,|\,\widehat{c}_h, \gamma_h) - \mathbb{P}_h^{\widehat{\mathcal{M}}(\pi^{1:H}),z}(z_{h+1}\,|\,\widehat{c}_h, \gamma_h)\right| \leq \theta_1 + 2AP_h\frac{\zeta_2}{\zeta_1} + AP_h\theta_2 + \frac{A^{2\widehat{L}}O^{\widehat{L}}\zeta_1}{\phi}.$$

Proof. It suffices to only consider $\pi \in \Pi^{\mathrm{det}}$. After some algebra, we can bound

$$\begin{aligned}
&\sum_{p_h, a_h, o_{h+1}}\left|\mathbb{P}_h^{\widetilde{\mathcal{M}}}(p_h, a_h, o_{h+1}\,|\,\widehat{c}_h, \gamma_h) - \mathbb{P}_h^{\widehat{\mathcal{M}}}(p_h, a_h, o_{h+1}\,|\,\widehat{c}_h, \gamma_h)\right| \\
&\leq \left\|\mathbb{P}_h^{\pi^h_{1:h-1},\mathcal{G}}(\cdot\,|\,\widehat{c}_h) - \mathbb{P}_h^{\widehat{\mathcal{M}}}(\cdot\,|\,\widehat{c}_h)\right\|_1 \\
&\quad + \sum_{p_h:\mathbb{P}_h^{\pi^h_{1:h-1},\mathcal{G}}(p_h|\widehat{c}_h)\geq\frac{\zeta_2}{\zeta_1}} \mathbb{P}_h^{\pi^h_{1:h-1},\mathcal{G}}(p_h\,|\,\widehat{c}_h)\sum_{a_h}\left\|\mathbb{P}_h^{\pi^h_{1:h-1},\mathcal{G}}(\cdot\,|\,\widehat{c}_h, p_h, a_h) - \mathbb{P}_h^{\widehat{\mathcal{M}}}(\cdot\,|\,\widehat{c}_h, p_h, a_h)\right\|_1 \\
&\quad + \sum_{p_h:\mathbb{P}_h^{\pi^h_{1:h-1},\mathcal{G}}(p_h|\widehat{c}_h)\leq\frac{\zeta_2}{\zeta_1}} \mathbb{P}_h^{\pi^h_{1:h-1},\mathcal{G}}(p_h\,|\,\widehat{c}_h)\sum_{a_h}\left\|\mathbb{P}_h^{\pi^h_{1:h-1},\mathcal{G}}(\cdot\,|\,\widehat{c}_h, p_h, a_h) - \mathbb{P}_h^{\widehat{\mathcal{M}}}(\cdot\,|\,\widehat{c}_h, p_h, a_h)\right\|_1,
\end{aligned}$$

where under the event $\mathcal{E}_1$, the first term can be bounded by the first item of Lemma 4. For the second term, since $\mathbb{P}_h^{\pi^h_{1:h-1},\mathcal{G}}(p_h\,|\,\widehat{c}_h) \geq \frac{\zeta_2}{\zeta_1}$, it implies that $\mathbb{P}_h^{\pi^h_{1:h-1},\mathcal{G}}(\widehat{c}_h, p_h) \geq \zeta_2$ together with the pre-condition that $\mathbb{P}_h^{\pi^h_{1:h-1},\mathcal{G}}(\widehat{c}_h) \geq \zeta_1$. This allows us to apply the second item of Lemma 4 to bound the corresponding $\left\|\mathbb{P}_h^{\pi^h_{1:h-1},\mathcal{G}}(\cdot\,|\,\widehat{c}_h, p_h, a_h) - \mathbb{P}_h^{\widehat{\mathcal{M}}}(\cdot\,|\,\widehat{c}_h, p_h, a_h)\right\|_1$. For the third term, we directly bound $\left\|\mathbb{P}_h^{\pi^h_{1:h-1},\mathcal{G}}(\cdot\,|\,\widehat{c}_h, p_h, a_h) - \mathbb{P}_h^{\widehat{\mathcal{M}}}(\cdot\,|\,\widehat{c}_h, p_h, a_h)\right\|_1$ by 2. Combining them together, we can conclude

$$\sum_{p_h, a_h, o_{h+1}}\left|\mathbb{P}_h^{\widetilde{\mathcal{M}}}(p_h, a_h, o_{h+1}\,|\,\widehat{c}_h, \gamma_h) - \mathbb{P}_h^{\widehat{\mathcal{M}}}(p_h, a_h, o_{h+1}\,|\,\widehat{c}_h, \gamma_h)\right| \leq \theta_1 + 2AP_h\frac{\zeta_2}{\zeta_1} + AP_h\theta_2.$$

Noticing that after marginalization, the total variation distance will not increase, we proved our lemma. $\square$

Until now, we have handled those $\widehat{c}_h$ such that $\mathbb{P}_h^{\pi^h_{1:h-1}}(\widehat{c}_h) \geq \zeta_1$. For those less visited $\widehat{c}_h$, we managed to bound the probability of visiting them by the probability of visiting certain less-explored states at step $h-\widehat{L}$, specifically $\sum_{s_{h-L}:\mathbb{P}_{h-L}^{\pi^h,\mathcal{G}}(s_{h-L})\leq\phi}\mathbb{P}_{h-L}^{\pi,\mathcal{G}}(s_{h-L})$ in Lemma 12. This finally corresponds to the error term $\epsilon_{apx}(\pi^{1:H})$ defined in Theorem 4 (c.f. the concrete expression in Equation (D.1)). By a triangle inequality, we can evaluate $\epsilon_z(\widehat{\mathcal{M}}(\pi^{1:H}))$ through the intermediate model $\widetilde{\mathcal{M}}(\pi^{1:H})$, leading to the final optimality guarantee for planning in $\widehat{\mathcal{M}}(\pi^{1:H})$ with the help of Theorem 2. Meanwhile, the sample complexity is simply $H \times N_0$, thus proving the sample complexity guarantee in Theorem 4.

6.4 Proof Outline of Theorem 5

Note that Theorem 4 characterizes the sample complexity for learning an equilibrium for $\mathcal{G}$ from the model $\widetilde{\mathcal{M}}(\pi^{1:H})$ with approximation errors depending on $\pi^{1:H}$. Therefore, to obtain the final guarantees, one needs to find certain policies $\pi^{1:H}$ to control the corresponding errors in Theorem 4, i.e., $\epsilon_r, \epsilon_z, \epsilon_{apx}$. Note that we have evaluated ϵ_{apx} above. For ϵ_r, ϵ_z, similar to the proof for Theorem 3, we take the one-step delayed sharing case as an example.

Step 1: Evaluating $\epsilon_z(\widetilde{\mathcal{M}}(\pi^{1:H}))$ and $\epsilon_r(\widetilde{\mathcal{M}}(\pi^{1:H}))$. We also use the finite-memory truncation as the compression as before. For any $\pi^{1:H}$, it is direct to verify that

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

where $\widetilde{\boldsymbol{b}}_h^{\pi^h}$ denotes the approximate belief state, where one ignores the history before step $h-L$ and performs the belief update using the Bayes rule after it, from the prior distribution of $\mathbb{P}_{h-L}^{\pi^h,\mathcal{G}}(s_{h-L})$(c.f. detailed definition in §E.6). If $L \geq C\gamma^{-4}\log(\frac{1}{\epsilon\phi})$, it holds that

$$\begin{aligned}
\epsilon_z(\widetilde{\mathcal{M}}(\pi^{1:H})) &= \max_h \max_{\pi\in\Pi^{\mathrm{det}},\gamma_h} \mathbb{E}_\pi^{\mathcal{G}} \left\| \mathbb{P}_h^{\mathcal{G}}(\cdot \,|\, c_h, \gamma_h) - \mathbb{P}_h^{\widetilde{\mathcal{M}},z}(\cdot \,|\, \widehat{c}_h, \gamma_h) \right\|_1 \\
&\leq \max_h \max_{\pi\in\Pi^{\mathrm{det}},\gamma_h} \mathbb{E}_\pi^{\mathcal{G}} \left\| \boldsymbol{b}_h(a_{1:h-1}, o_{1:h-1}) - \widetilde{\boldsymbol{b}}_h^{\pi^h}(a_{h-L:h-1}, o_{h-L+1:h-1}) \right\|_1 \\
&\leq \epsilon + \max_h \max_{\pi\in\Pi^{\mathrm{det}}} \mathbf{1}[h > L] \cdot 6 \cdot \sum_{s_{h-L}:\mathbb{P}_{h-L}^{\pi^h,\mathcal{G}}(s_{h-L})\leq\phi} \mathbb{P}_{h-L}^{\pi,\mathcal{G}}(s_{h-L}),
\end{aligned} \tag{6.3}$$

where the last step is again adapted from the belief contraction results under γ-observability of Golowich et al. (2022b) (c.f. Lemma 15). $\epsilon_r(\widetilde{\mathcal{M}}(\pi^{1:H}))$ can be evaluated similarly.

Step 2: Minimizing the visitation probability of less-explored states with Barycentric Spanner. Now to control ϵ_z, ϵ_r and ϵ_{apx}, it suffices to control the key quantity, $\sum_{s_{h-L}:\mathbb{P}_{h-L}^{\pi^h,\mathcal{G}}(s_{h-L})\leq\phi}\mathbb{P}_{h-L}^{\pi,\mathcal{G}}(s_{h-L})$. In other words, $\pi^{1:H}$ should be *exploratory* enough in the sense that the actual states are visited often enough. It turns out that finding such exploratory policies to minimize this error term can be achieved by the Barycentric-spanner-based techniques (Awerbuch and Kleinberg, 2008), as also adopted by Golowich et al. (2022a), using quasi-polynomial sample and computational complexities (c.f. Lemma 16). By choosing the parameters ζ_1, ζ_2, θ_1, θ_2, and ϕ properly, we proved Theorem 5.

6.5 Proof Outline of Theorem 6

Correctness of the algorithmic framework. The correctness of our framework follows similarly from the proof of Theorem 2. Combining the fact that $\widehat{\pi}^\star$ is an optimal policy of $\mathcal{M}$ and Lemma 2, $\widehat{\pi}^\star$ is also an approximate optimal policy of $\mathcal{G}$.

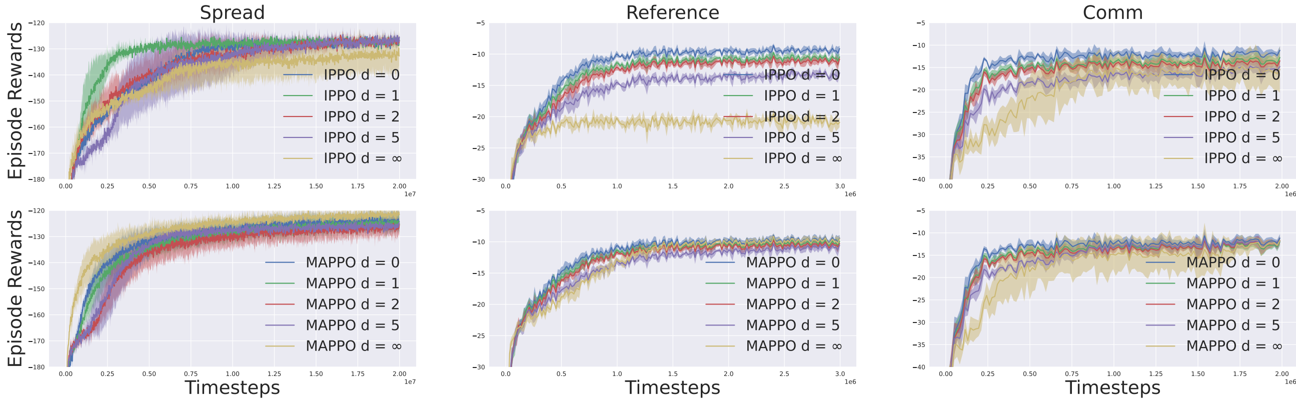

Figure 1: Performance of MAPPO and IPPO in various delayed-sharing settings.

Computation analysis. As we mentioned in §5, the key of extending our framework to team-optimum-finding in Dec-POMDPs is to implement Equation (5.1) in a computationally tractable way for each $h \in [H]$. Here we briefly outline how each of the three assumptions can circumvent the hardness in Proposition 2. **Condition 1.** In Proposition 11, we show that, the Q-value function can be linearly decomposed into n functions, i.e., $Q_h^{\star,\mathcal{M}}(\widehat{c}_h, \gamma_h) = \sum_{j\in[n]} U_{j,h}(\widehat{c}_h, \gamma_{j,h})$, for some functions $\{U_{j,h}\}_{j\in[n]}$, for any $\widehat{c}_h \in \widehat{\mathcal{C}}_h$, $\gamma_h \in \Gamma_h$. Therefore, Equation (5.1) can be solved tractably since each $U_{j,h}$ is indeed a *linear* function of $\gamma_{j,h}$, with the *concatenation* of simplexes being the constraint. **Condition 2.** With such a nested structure, Equation (5.1) can indeed be solved by a *dynamic programming over the agents*. We consider the following POMDP $\widehat{\mathcal{P}}(n)$ with the horizon-length being the number of the agents n. The initial state $x_1 = (s_h, p_h) \sim \mathbb{P}_h^{\mathcal{M},c}(s_h, p_h \mid \widehat{c}_h)$. At each step $j \in [n]$ of this POMDP, the observation is $y_j = p_{j,h}$, the agent j takes the action $a_j \in \mathcal{A}_j$, and the next state transitions to $x_{j+1} = (x_j, a_j)$. Note that the reward is non-zero only at the last step n, where $\widehat{r}_n(x_n, a_n) = \mathbb{E}_{s_{h+1}\sim\mathbb{T}_h(\cdot\mid s_h, a_{1:n}), o_{h+1}\sim\mathbb{O}_{h+1}(\cdot\mid s_{h+1})}[r_h(s_h, a_{1:n}) + V_{h+1}^{\star,\mathcal{M}}(\widehat{c}_{h+1})]$. Based on such a POMDP perspective, we develop Algorithm 10, where the first for-loop is a standard *backward procedure* of value iteration for the POMDP $\widehat{\mathcal{P}}(n)$ constructed above to compute its optimal policy $u_{1:n}^\star$. The second for-loop performs a *forward procedure* of translating $u_{1:n}^\star$ into $\gamma_{1:n,h}^\star$, where $\gamma_{i,h}^\star \in \Gamma_{i,h}$ for each $i \in [n]$ now belongs to the prescription space we hope to optimize over in Equation (5.1). Note that throughout, we regard the number of agents n, i.e., the time horizon of $\widehat{\mathcal{P}}(n)$ as a (small) constant. Hence, the time complexity of such a dynamic programming for finding the exact optimal policy of $\widehat{\mathcal{P}}(n)$ is indeed $\texttt{poly}(S, A, P_h)$. **Condition 3.** Due to the factorized structures, we show in Proposition 13 that the Q-value can be decoupled into n terms, such that there exist n functions $\{F_{i,h}\}_{i\in[n]}$ such that $Q_h^{\star,\mathcal{M}}(\widehat{c}_h, \{\gamma_{i,h}\}_{i\in[n]}) = \sum_{i\in[n]} F_{i,h}(\widehat{c}_{i,h}, \gamma_{i,h})$. Therefore, the maximization over the *joint* $\{\gamma_{i,h}\}_{i\in[n]}$ in Equation (5.1) is equivalent to the *individual* maximization over each $\gamma_{i,h}$ for $F_{i,h}$, $i \in [n]$, which is again a linear program as we argued before. Thus, Equation (5.1) can be also solved with time complexity $\texttt{poly}(S, A, P_h)$.

7 Experimental Results

For the experiments, we shall both investigate the benefits of *information sharing* as we considered in various empirical MARL environments, and validate the implementability and performance of our proposed approaches on several modest-scale examples.

	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 of the rewards using our methods, compared with using the methods of FM-E and RNN-E in Mao et al. (2020).

Information sharing improves performance. We mainly consider three cooperative tasks, the *physical deception* (*Spread*), the *simple reference* (*Reference*), and the *cooperative communication* (*Comm*) in the popular deep MARL benchmarks, multi-agent particle-world environment (MPE) (Lowe et al., 2017). We train both the popular centralized-training algorithm MAPPO (Yu et al., 2021) and the decentralized-training algorithm IPPO (Yu et al., 2021) with different information-sharing mechanisms by varying the delay from 0 to ∞. The rewards during training are shown in Figure 1. It is seen that in all domains (except MAPPO on Spread) with either training paradigms, smaller delays, which correspond to the case of more information sharing, will lead to faster convergence, higher final performance, and reduced training variance.

Validating implementability and performance. To further validate the tractability of our approaches, we test our learning algorithm on two popular and modest-scale partially observable benchmarks Dectiger (Nair et al., 2003) and Boxpushing (Seuken and Zilberstein, 2012). We compare our approaches with FM-E and RNN-E, which are also common information-based approaches developed 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.

8 Concluding Remarks

In this paper, we studied provable multi-agent RL in partially observable environments, with both statistical and computational (quasi-)efficiencies. The key to our results is to identify the value of *information sharing*, a common practice in empirical MARL and a standard phenomenon in many multi-agent control systems, in algorithm design and computation/sample efficiency analysis. We hope our study may open up the possibilities of leveraging and even designing different *information structures*, for developing both statistically and computationally efficient partially observable MARL algorithms. One open problem and future direction is to develop a fully *decentralized* algorithm and overcome *the curse of multiagents*, such that the sample and computation complexities do not grow exponentially with the number of agents. Another interesting direction is to identify the combination of certain information-sharing structures and observability assumptions for more efficient (e.g., polynomial) sample and computation complexity results.

Acknowledgement

The authors would like to thank the anonymous reviewers from ICML 2023 for their helpful comments. The authors would also like to thank Noah Golowich and Serdar Yüksel for the valuable feedback and discussions. K.Z. acknowledges support from Simons-Berkeley Research Fellowship, Northrop Grumman – Maryland Seed Grant Program, the U.S. Army Research Laboratory and the U.S. Army Research Office under grant number W911NF-24-1-0085, and NSF CAREER Award 2443704.

References

Altabaa, A. and Yang, Z. (2024). On the role of information structure in reinforcement learning for partially-observable sequential teams and games. *arXiv preprint arXiv:2403.00993*.

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

Aumann, R. J., Maschler, M. and Stearns, R. E. (1995). *Repeated games with incomplete information*. MIT press.

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

Azizzadenesheli, K., Lazaric, A. and Anandkumar, A. (2016). Reinforcement learning of POMDPs using spectral methods. In *Conference on Learning Theory*. PMLR.

Bai, Y. and Jin, C. (2020). Provable self-play algorithms for competitive reinforcement learning. In *International Conference on Machine Learning*.

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

Behn, R. and Ho, Y.-C. (1968). On a class of linear stochastic differential games. *IEEE Transactions on Automatic Control*, **13** 227–240.

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

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

Blum, A. and Mansour, Y. (2007). Learning, regret minimization, and equilibria. *Algorithmic Game Theory* 79–102.

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

Cai, Y., Liu, X., Oikonomou, A. and Zhang, K. (2024). Provable partially observable reinforcement learning with privileged information. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

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

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

Chen, Z., Zhang, K., Mazumdar, E., Ozdaglar, A. and Wierman, A. (2023). A finite-sample analysis of payoff-based independent learning in zero-sum stochastic games. *arXiv preprint arXiv:2303.03100*.

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

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

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

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

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

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

Emery-Montemerlo, R., Gordon, G., Schneider, J. and Thrun, S. (2004). 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.* IEEE.

Even-Dar, E., Kakade, S. M. and Mansour, Y. (2007). The value of observation for monitoring dynamic systems. In *IJCAI*.

Farina, G., Anagnostides, I., Luo, H., Lee, C.-W., Kroer, C. and Sandholm, T. (2022). Near-optimal no-regret learning dynamics for general convex games. *Advances in Neural Information Processing Systems*, **35** 39076–39089.

Filar, J. and Vrieze, K. (2012). *Competitive Markov Decision Processes*. Springer Science & Business Media.

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

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

Golowich, N., Moitra, A. and Rohatgi, D. (2023). Planning and learning in partially observable systems via filter stability. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*.

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

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

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

Hansen, E. A., Bernstein, D. S. and Zilberstein, S. (2004). Dynamic programming for partially observable stochastic games. In *AAAI*, vol. 4.

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

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

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

Jin, C., Liu, Q., Wang, Y. and Yu, T. (2024). V-learning – a simple, efficient, decentralized algorithm for multiagent reinforcement learning. *Mathematics of Operations Research*, **49** 2295–2322.

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

Kao, H., Wei, C.-Y. and Subramanian, V. (2022). Decentralized cooperative reinforcement learning with hierarchical information structure. In *International Conference on Algorithmic Learning Theory*. PMLR.

Kara, A. D. (2022). Near optimality of finite memory feedback policies in partially observed Markov decision processes. *The Journal of Machine Learning Research*, **23** 437–482.

Kara, A. D. and Yüksel, S. (2022). Convergence of finite memory Q learning for POMDPs and near optimality of learned policies under filter stability. *Mathematics of Operations Research*.

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

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

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

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

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

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

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

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

Liu, X. and Zhang, K. (2023). Partially observable multi-agent rl with (quasi-) efficiency: the blessing of information sharing. In *International Conference on Machine Learning*. PMLR.

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

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

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

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

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

Milgrom, P. and Roberts, J. (1987). Informational asymmetries, strategic behavior, and industrial organization. *The American Economic Review*, **77** 184–193.

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

Nair, R., Tambe, M., Yokoo, M., Pynadath, D. and Marsella, S. (2003). Taming decentralized pomdps: Towards efficient policy computation for multiagent settings. In *IJCAI*, vol. 3.

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

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

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

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

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

Pathak, A., Pucha, H., Zhang, Y., Hu, Y. C. and Mao, Z. M. (2008). 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*. Springer.

Peralez, J., Delage, A., Buffet, O. and Dibangoye, J. S. (2024). Solving hierarchical information-sharing dec-pomdps: An extensive-form game approach. *arXiv preprint arXiv:2402.02954*.

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

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

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

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

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

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

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

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

Sinha, A. and Mahajan, A. (2023). Asymmetric actor-critic with approximate information state. In *2023 62nd IEEE Conference on Decision and Control (CDC)*. IEEE.

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

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

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

Tang, D., Subramanian, V. and Teneketzis, D. (2024). Information compression in dynamic games. *arXiv preprint arXiv:2407.12318*.

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

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

Wei, C.-Y., Lee, C.-W., Zhang, M. and Luo, H. (2021). Last-iterate convergence of decentralized optimistic gradient descent/ascent in infinite-horizon competitive Markov games. *arXiv preprint arXiv:2102.04540*.

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

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

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

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

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

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

Zhang, K., Zhang, X., Hu, B. and Basar, T. (2021b). 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.

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

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

Appendices

A Additional Definitions

A.1 Belief states

In partially observable environments, 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 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 11 (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 $1 \leq h \leq H$ and any action-observation sequence $(a_{1:h-1}, o_{1:h})$, we inductively define the belief state:

$$\boldsymbol{b}_{h+1}(a_{1:h}, o_{1:h}) = \mathbb{T}_h(a_h) \cdot \boldsymbol{b}_h(a_{1:h-1}, o_{1:h}),$$

$$\boldsymbol{b}_h(a_{1:h-1}, o_{1:h}) = B_h(\boldsymbol{b}_h(a_{1:h-1}, o_{1:h-1}); o_h).$$

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

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

We define the approximate belief update using the most recent L-step history. For $1 \leq h \leq H$, 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 $1 \leq h - L < h \leq H$ and any action-observation sequence $(a_{h-L:h-1}, o_{h-L+1:h})$, we inductively define

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

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

For the remainder of our paper, we shall use the important initialization for the approximate belief, which are defined as $\boldsymbol{b}_h'(\cdot) := \boldsymbol{b}_h^{\mathrm{apx},\mathcal{G}}(\cdot; \mathrm{Unif}(\mathcal{S}))$.

A.2 Additional definitions of value functions and policies

In Definition 1, we have defined value functions in $\mathcal{G}$. Similar to the fully-observable settings (MDPs and stochastic games), we can also extend such a definition to the prescription-value function, which corresponds to the action-value function in the fully-observable settings.

Definition 12 (Prescription-value function with information sharing)**.** *At step $h \in [H]$, given the common information c_h, joint policies $\pi = \{\pi_i\}_{i=1}^n \in \Pi$, and prescriptions $\{\gamma_{i,h}\}_{i=1}^n \in \Gamma_h$, the* prescription-value function conditioned on the common information and joint prescription *of the i^{th} agent is defined as:*

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

where prescription $\gamma_{i,h} \in \Gamma_{i,h}$ replaces the partial function $\pi_{i,h}(\cdot \,|\, \omega_{i,h}, c_h, \cdot)$ in the value function.

With the expected approximate common information model $\mathcal{M}$ given in Definition 7, we can define the value function and policy under $\mathcal{M}$ accordingly as follows.

Definition 13 (Value function and policy under $\mathcal{M}$)**.** *Given an expected 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*

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

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*

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

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 (A.1), respectively.*

B Information Sharing in Applications

The 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 usually 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.

Example 1 (One-step delayed sharing)**.** *At any step $h \in [H+1]$, the common and private information are given as $c_h = \{o_{1:h-1}, a_{1:h-1}\}$ 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 the 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 the team setting to be focused on later (c.f. §5), we additionally assume, for this example, that the reward function has an additive structure, i.e., $r_h(s_h, a_h) = \sum_{j\in[n]} r_{j,h}(s_h, a_{j,h})$ for some functions $\{r_{j,h}\}_{j\in[n]}$. The information structure is given as $c_h = \{o_{1,1:h}, o_{2,1:h-d}, a_{1,1:h-1}\}$, $p_{1,h} = \emptyset$, $p_{2,h} = \{o_{2,h-d+1:h}\}$, 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. We will regard d as a constant throughout. 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 the 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 $c_h = \{o_{1:h}, a_{1:h-1}\}$ and $p_{i,h} = \emptyset$. We will also denote this structure as* fully sharing *hereafter.*

Example 4 (Information sharing with one-directional-one-step delay)**.** *Similar to the previous cases, we also assume there are 2 players for ease of exposition, and the case can be generalized to multi-player cases straightforwardly. Similar to the one-step delay case, we consider the situation where all observations of player 1 are available to player 2, while the observations of player 2 are available to player 1 with one-step delay. All the past actions are available to both players. That is, in this case, $c_h = \{o_{1,1:h}, o_{2,2:h-1}, a_{1:h-1}\}$, and player 1 has no private information, i.e., $p_{1,h} = \emptyset$, 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. For the reward, we assume it still depends on the action of one agent to avoid trivial solutions at each step h. An example of this case is the information structure where controllers share their observations with a general delay of $d \geq 1$ time steps. In this case, the common information is $c_h = \{o_{1:h-d}\}$ and the private information is $p_{i,h} = \{o_{i,h-d+1:h}\}$. Such information structures can be used to model repeated games with incomplete information (Aumann et al., 1995).*

C Collection of Algorithm Pseudocodes

Here we collect both our planning and learning algorithms as in Algorithms 1, 2, 3, 4, 5, 6, 7, 8, 9.

D Full Versions of the Results

D.1 Planning

Now we state the full version of Theorem 3 regarding the instantiations of Theorem 2.

Theorem 7. *Fix $\epsilon > 0$. Suppose there exists an (ϵ_r, ϵ_z)-expected-approximate common information model $\mathcal{M}$ consistent with some given approximate belief $\{\mathbb{P}_h^{\mathcal{M},c}(s_h, p_h\,|\,\widehat{c}_h)\}_{h\in[H]}$ for the POSG $\mathcal{G}$ under Assumptions 1 and 3 such that $\max\{\epsilon_z(\mathcal{M}), \epsilon_r(\mathcal{M})\} \leq \mathcal{O}(\epsilon)$ and $\max_h \widehat{C}_h P_h$ is quasi-polynomial of the problem instance size, then there exists a quasi-polynomial time algorithm that can compute an ϵ-NE if $\mathcal{G}$ is zero-sum or cooperative, and an ϵ-CE/CCE if $\mathcal{G}$.*

In particular, under Assumption 2, examples in Appendix B satisfy all such conditions. Therefore, 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 in Assumption 2:

- ***One-step delayed information sharing:*** $(AO)^{C\gamma^{-4}\log\frac{SH}{\epsilon}}$ *for some universal constant $C > 0$.*

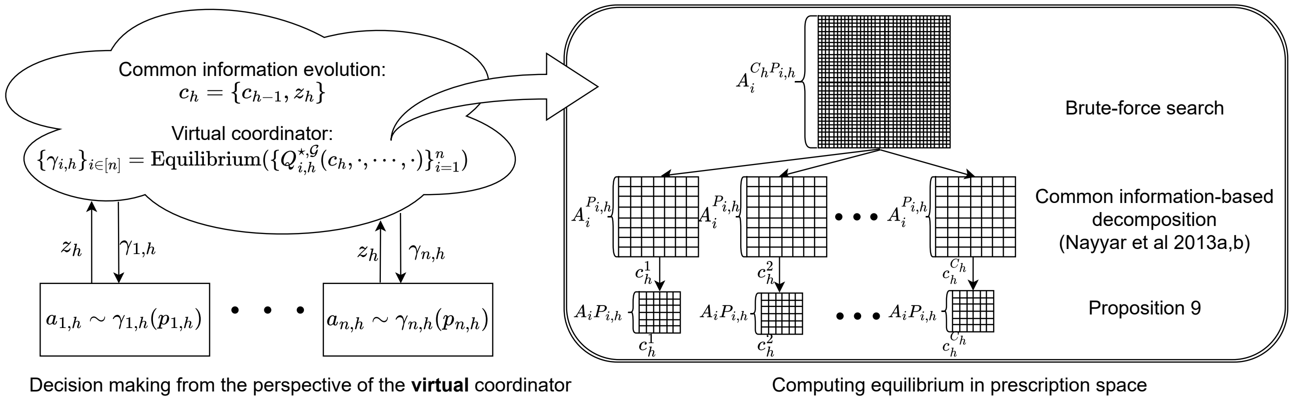

Figure 2: An overview of our algorithmic framework. The left part of the figure shows that there is a virtual coordinator collecting the information shared among agents. Based on the common information c_h, it will compute an equilibrium in the prescription space and assign it to all the agents. The 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) proposed the common information-based decomposition, and solve C_h number of games of smaller size. However, in the Dec-POMDP setting, Nayyar et al. (2013b) treated each deterministic prescription as an action and the size of each sub-problem will be $A_i^{P_{i,h}} = 8$. Furthermore, Proposition 8 shows that we can reformulate each sub-problem as a game whose payoff is multi-linear with respect to each agent's prescription, and whose dimensionality is $A_i P_{i,h} = 6$.

- ***State controlled by one controller with asymmetric** $d = \text{poly}(\log H)$**-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** $d = \text{poly}(\log H)$**-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.2 Learning

Here we state the full version of Theorem 4 regarding the sample efficiency of learning and approximate common information model.

Theorem 8. *Suppose the POSG $\mathcal{G}$ satisfies Assumptions 1 and 3. Given any compression functions of common information,* $\text{Compress}_h : \mathcal{C}_h \to \widehat{\mathcal{C}}_h$ *for* $h \in [H+1]$*, we can compute* $\widehat{L}$ *as defined in Definition 10. Then, given any H policies* $\pi^{1:H}$*, where* $\pi^h \in \Delta(\Pi^{\text{det}})$*,* $\pi^h_{h-\widehat{L}:h} = \text{Unif}(\mathcal{A})$ *for* $h \in [H]$*, we can construct a policy-dependent expected approximate common information model* $\widetilde{\mathcal{M}}(\pi^{1:H})$*, whose compression functions are* $\{\text{Compress}_h\}_{h\in[H+1]}$*. We write* $\epsilon_r(\pi^{1:H}) := \epsilon_r(\widetilde{\mathcal{M}}(\pi^{1:H}))$ *and* $\epsilon_z(\pi^{1:H}) := \epsilon_z(\widetilde{\mathcal{M}}(\pi^{1:H}))$ *for short. 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*

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

$$Q^{\star,\mathcal{G}}_{i,h}(c_h, \gamma_{1,h}, \cdots, \gamma_{n,h}) := \mathbb{E}_{s_h, p_h \sim \mathbb{P}^{\mathcal{G}}_h(\cdot,\cdot\,|\,c_h)} \mathbb{E}_{\{a_{j,h} \sim \gamma_{j,h}(\cdot\,|\,p_{j,h})\}_{j\in[n]}} \mathbb{E}_{o_{h+1} \sim \mathbb{O}^{\top}_{h+1}\mathbb{T}_h(\cdot\,|\,s_h, a_h)} \Big[r_{i,h}(s_h, a_h) + V^{\star,\mathcal{G}}_{i,h+1}(c_{h+1}) \Big]$$

$$\Big\{ \pi^{\star}_{1,h}(\cdot\,|\,\cdot, c_h, \cdot), \cdots, \pi^{\star}_{n,h}(\cdot\,|\,\cdot, c_h, \cdot) \Big\} \leftarrow \text{NE/CE/CCE}(\{Q^{\star,\mathcal{G}}_{i,h}(c_h, \cdot, \cdots, \cdot)\}_{i=1}^{n}, \epsilon_e)$$

        // we refer the implementation to Appendix E.2
        for each $i \in [n]$ **do**

$$V^{\star,\mathcal{G}}_{i,h}(c_h) \leftarrow \mathbb{E}_{\{\omega_{j,h}\}_{j\in[n]}} \mathbb{E}^{\mathcal{G}} \Big[r_{i,h}(s_h, a_h) + V^{\star,\mathcal{G}}_{i,h+1}(c_{h+1}) \,|\, c_h, \{\pi^{\star}_{j,h}(\cdot\,|\,\omega_{j,h}, c_h, \cdot)\}_{j\in[n]} \Big]$$

        end for
    end for
end for
return $\pi^{\star}$

Algorithm 2 $\mathsf{BR}(\mathcal{G}, \pi, i, \epsilon_e)$: ϵ_e-approximate **B**est **R**esponse for agent i under true model $\mathcal{G}$

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

$$Q^{\star,\mathcal{G}}_{i,h}(c_h, \gamma_{1,h}, \cdots, \gamma_{n,h}) := \mathbb{E}_{s_h, p_h \sim \mathbb{P}^{\mathcal{G}}_h(\cdot,\cdot\,|\,c_h)} \mathbb{E}_{\{a_{j,h} \sim \gamma_{j,h}(\cdot\,|\,p_{j,h})\}_{j\in[n]}} \mathbb{E}_{o_{h+1} \sim \mathbb{O}^{\top}_{h+1}\mathbb{T}_h(\cdot\,|\,s_h, a_h)} \Big[r_{i,h}(s_h, a_h) + V^{\star,\mathcal{G}}_{i,h+1}(c_{h+1}) \Big]$$

$$\pi^{\star}_{i,h}(\cdot\,|\,\cdot, c_h, \cdot) \leftarrow \text{NE/CE/CCE-BR}(Q^{\star,\mathcal{G}}_{i,h}(c_h, \cdot, \cdots, \cdot), \{\pi_{j,h}(\cdot\,|\,\cdot, c_h, \cdot)\}_{j\in[n]}, i, \epsilon_e)$$

        // we refer the implementation to Appendix E.2

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

    end for
end for
return $\pi^{\star}_i$

Algorithm 3 `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**
  $V_{i,H+1}^{\star,\mathcal{M}}(\widehat{c}_{H+1}) \leftarrow 0$
end for
for $h = H, \cdots, 1$ **do**
  for each $\widehat{c}_h$ **do**
    Define $Q_{i,h}^{\star,\mathcal{M}}(\widehat{c}_h, \gamma_{1,h}, \cdots, \gamma_{n,h}) := \widehat{r}_{i,h}^{\mathcal{M}}(\widehat{c}_h, \gamma_h) + \mathbb{E}^{\mathcal{M}}\left[V_{i,h+1}^{\star,\mathcal{M}}(\widehat{c}_{h+1}) \mid \widehat{c}_h, \{\gamma_{j,h}\}_{j\in[n]}\right]$ for any $i \in [n]$
    if computing the equilibrium **then**

$$\left\{\widehat{\pi}_{1,h}^{\star}(\cdot \mid \cdot, \widehat{c}_h, \cdot), \cdots, \widehat{\pi}_{n,h}^{\star}(\cdot \mid \cdot, \widehat{c}_h, \cdot)\right\} \leftarrow \text{NE/CE/CCE}(\{Q_{i,h}^{\star,\mathcal{M}}(\widehat{c}_h, \cdot, \cdots, \cdot)\}_{i=1}^{n}, \epsilon_e)$$

    // we refer the implementation to Appendix E.2
    else if computing the team-optimum **then**

$$\left\{\widehat{\pi}_{1,h}^{\star}(\cdot \mid \widehat{c}_h, \cdot), \cdots, \widehat{\pi}_{n,h}^{\star}(\cdot \mid \widehat{c}_h, \cdot)\right\} \leftarrow \arg \max_{\left\{\gamma_{i,h} \in \Delta(\mathcal{A}_i)^{\mathcal{P}_{i,h}}\right\}_{i\in[n]}} (Q_{1,h}^{\star,\mathcal{M}}(\widehat{c}_h, \gamma_{1,h}, \cdots, \gamma_{n,h}))$$

    // we refer the implementation to Appendix E.7
    end if
    for each $i \in [n]$ **do**
      $V_{i,h}^{\star,\mathcal{M}}(\widehat{c}_h) \leftarrow \mathbb{E}_{\{\omega_{j,h}\}_{j\in[n]}}\left[\widehat{r}_{i,h}^{\mathcal{M}}(\widehat{c}_h, \{\widehat{\pi}_{j,h}^{\star}(\cdot \mid \omega_{j,h}, \widehat{c}_h, \cdot)\}_{j\in[n]}) + \mathbb{E}^{\mathcal{M}}[V_{i,h+1}^{\star,\mathcal{M}}(\widehat{c}_{h+1}) \mid \widehat{c}_h, \{\widehat{\pi}_{j,h}^{\star}(\cdot \mid \omega_{j,h}, \widehat{c}_h, \cdot)\}_{j\in[n]}]\right]$
    end for
  end for
end for
return $\widehat{\pi}^{\star}$

approximation error for estimating $\widetilde{\mathcal{M}}(\pi^{1:H})$ using samples under the policies $\pi^{1:H}$ as:

$$\epsilon_{apx}(\pi^{1:H}, \widehat{L}, \zeta_1, \zeta_2, \theta_1, \theta_2, \phi) = \theta_1 + 2A \max_h P_h \frac{\zeta_2}{\zeta_1} + A \max_h P_h \theta_2 + \frac{A^{2\widehat{L}} O^{\widehat{L}} \zeta_1}{\phi} + \max_h \max_{\pi \in \Pi^{\det}} \mathbf{1}[h > \widehat{L}] \cdot 2 \cdot d_{\mathcal{S},h-\widehat{L}}^{\pi,\mathcal{G}}\left(\mathcal{U}_{\phi,h-\widehat{L}}^{\mathcal{G}}(\pi^h)\right), \tag{D.1}$$

where for any policy $\pi' \in \Delta(\Pi^{\det}), h \in [H]$, we define $d_{\mathcal{S},h}^{\pi',\mathcal{G}}(s) := \mathbb{P}_h^{\pi',\mathcal{G}}(s_h = s)$, $d_{\mathcal{S},h}^{\pi',\mathcal{G}}(A) := \sum_{s\in A} d_{\mathcal{S},h}^{\pi',\mathcal{G}}(s)$ for any $A \subseteq \mathcal{S}$, $\mathcal{U}_{\phi,h}^{\mathcal{G}}(\pi') := \{s \in \mathcal{S} : d_{\mathcal{S},h}^{\pi',\mathcal{G}}(s) < \phi\}$, representing under-explored states under the policy π'. Then, Algorithm 5 can learn an model $\widehat{\mathcal{M}}(\pi^{1:H})$ with the sample complexity

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

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

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

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 $Q^{\star,\mathcal{M}}_{i,h}(\widehat{c}_h, \gamma_{1,h}, \cdots, \gamma_{n,h}) := \widehat{r}^{\mathcal{M}}_{i,h}(\widehat{c}_h, \gamma_h) + \mathbb{E}^{\mathcal{M}}\Big[V^{\star,\mathcal{M}}_{i,h+1}(\widehat{c}_{h+1}) \mid \widehat{c}_h, \{\gamma_{j,h}\}_{j\in[n]}\Big]$ for any $i \in [n]$

$$\widehat{\pi}^{\star}_{i,h}(\cdot \mid \cdot, \widehat{c}_h, \cdot) \leftarrow \text{NE/CE/CCE-BR}(Q^{\star,\mathcal{M}}_{i,h}(\widehat{c}_h, \cdot, \cdots, \cdot), \{\widehat{\pi}_{j,h}(\cdot \mid \cdot, \widehat{c}_h, \cdot)\}_{j\in[n]}, i, \epsilon_e)$$

  // we refer the implementation to Appendix E.2
    $V^{\star,\mathcal{M}}_{i,h}(\widehat{c}_h) \leftarrow \mathbb{E}_{\{\omega_{j,h}\}_{j\in[n]}}\Big[\widehat{r}^{\mathcal{M}}_{i,h}(\widehat{c}_h, \{\widehat{\pi}^{\star}_{i,h}(\cdot \mid \omega_{i,h}, \widehat{c}_h, \cdot), \widehat{\pi}_{-i,h}(\cdot \mid \omega_{-i,h}, \widehat{c}_h, \cdot)\}) + \mathbb{E}^{\mathcal{M}}[V^{\star,\mathcal{M}}_{i,h+1}(\widehat{c}_{h+1}) \mid \widehat{c}_h, \{\widehat{\pi}^{\star}_{i,h}(\cdot \mid \omega_{i,h}, \widehat{c}_h, \cdot), \widehat{\pi}_{-i,h}(\cdot \mid \omega_{-i,h}, \widehat{c}_h, \cdot)\}]\Big]$
  end for
end for
return $\widehat{\pi}^{\star}_i$

$i \in [n]$:

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

Under such a high probability event, the policy output of Algorithm 3 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

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

We state the full version of Theorem 5 regarding the instantiation of Theorem 4 in the following.

Theorem 9. *Fix $\epsilon, \delta > 0$. Suppose the POSG $\mathcal{G}$ satisfies Assumptions 1 and 3. If there exist some compression functions of common information,* $\text{Compress}_h : \mathcal{C}_h \to \widehat{\mathcal{C}}_h$ *for $h \in [H+1]$, $\pi^{1:H}$, and $\widetilde{\mathcal{M}}(\pi^{1:H})$ satisfying the conditions in Theorem 4, and there exists some parameters $\delta_1, \theta_1, \theta_2, \zeta_1, \zeta_2 > 0$ for Algorithm 5, $\epsilon_e > 0$ for Algorithm 3, and some $\phi > 0$, such that*

$$\max\left\{\epsilon_z(\pi^{1:H}), \epsilon_r(\pi^{1:H}), \epsilon_{apx}(\pi^{1:H}, \widehat{L}, \zeta_1, \zeta_2, \theta_1, \theta_2, \phi)\right\} \leq \mathcal{O}(\epsilon)$$

and $N_0 = \text{poly}(\max_{h\in[H]} P_h, \max_{h\in[H]} \widehat{C}_h, H, A, O, \frac{1}{\zeta_1}, \frac{1}{\zeta_2}, \frac{1}{\theta_1}, \frac{1}{\theta_2}) \cdot \log\frac{1}{\delta_1}$ is quasi-polynomial of the problem instance size, then Algorithm 5, together with Algorithm 3, can 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 $1-\delta$, using quasi-polynomial time and samples, where $\widehat{L}$ is defined as in Definition 10.

In particular, under Assumption 2, examples in Appendix B satisfy such all such conditions. Then, there exists a multi-agent RL algorithm (Algorithm 9) that, with probability at least $1-\delta$, learns an 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 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 $C > 0$.*

Algorithm 5 $\texttt{LEE}(\pi^{1:H}, \{\widehat{\mathcal{C}}_h\}_{h\in[H+1]}, \{\widehat{\phi}_{h+1}\}_{h\in[H]}, \Gamma, \zeta_1, \zeta_2, \theta_1, \theta_2, \delta_1)$: **L**earning **E**mpirical **E**stimator $\widehat{\mathcal{M}}(\pi^{1:H})$ of $\widetilde{\mathcal{M}}(\pi^{1:H})$

Input: $\pi^{1:H}, \{\widehat{\mathcal{C}}_h\}_{h\in[H+1]}, \{\widehat{\phi}_{h+1}\}_{h\in[H]}, \Gamma, \zeta_1, \zeta_2, \theta_1, \theta_2, \delta_1$
for $1 \le h \le H$ **do**
    Define N_0 as in Equation (D.2).
    Draw N_0 independent trajectories by executing the policy π^h, and denote the k^{th} trajectory by $(a^k_{1:H-1}, o^k_{1:H}, r^k_{1:H})$, for $k \in [N_0]$, where N_0 is specified in Theorem 4.
    for each $\widehat{c}_h \in \widehat{\mathcal{C}}_h$ **do**
        Define $\varphi(p_h) := |\{k : \text{Compress}_h(f_h(a^k_{1:h-1}, o^k_{1:h})) = \widehat{c}_h, \text{ and } g_h(a^k_{1:h-1}, o^k_{1:h}) = p_h\}|$.
        Set $\mathbb{P}_h^{\widehat{\mathcal{M}}(\pi^{1:H})}(p_h \mid \widehat{c}_h) := \frac{\varphi(p_h)}{\sum_{p'_h} \varphi(p'_h)}$ 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 $\mathbb{P}_h^{\widehat{\mathcal{M}}(\pi^{1:H})}(o_{h+1} \mid \widehat{c}_h, p_h, a_h) := \frac{\psi(o_{h+1})}{\sum_{o'_{h+1}} \psi(o'_{h+1})}$ for all $o_{h+1} \in \mathcal{O}$.
        Define $\kappa(\widehat{c}_h, p_h, a_h) := \{k : \text{Compress}_h(f_h(a^k_{1:h-1}, o^k_{1:h})) = \widehat{c}_h, g_h(a^k_{1:h-1}, o^k_{1:h}) = p_h, a^k_h = a_h, \text{ and } o^k_{h+1} = o_{h+1}\}$.
        Set $\widehat{r}_{i,h}^{\widehat{\mathcal{M}}(\pi^{1:H})}(\widehat{c}_h, p_h, a_h) := \frac{\sum_{k\in\kappa(\widehat{c}_h, p_h, a_h)} r^k_{i,h}}{|\kappa(\widehat{c}_h, p_h, a_h)|}$ for all $i \in [n]$.
    end for
end for
Define for any $h \in [H]$, $\widehat{c}_h \in \widehat{\mathcal{C}}_h$, $\gamma_h \in \Gamma_h$, $o_{h+1} \in \mathcal{O}_{h+1}$, $z_{h+1} \in \mathcal{Z}_{h+1}$:

$$\mathbb{P}_h^{\widehat{\mathcal{M}}(\pi^{1:H}),z}(z_{h+1} \mid \widehat{c}_h, \gamma_h) \leftarrow \sum_{p_h, a_h, o_{h+1}} \mathbf{1}[\chi_{h+1}(p_h, a_h, o_{h+1}) = z_{h+1}] \times \mathbb{P}_h^{\widehat{\mathcal{M}}(\pi^{1:H})}(p_h \mid \widehat{c}_h)\gamma_h(a_h \mid p_h)\mathbb{P}_h^{\widehat{\mathcal{M}}(\pi^{1:H})}(o_{h+1} \mid \widehat{c}_h, p_h, a_h)$$

$$\widehat{r}_{i,h}^{\widehat{\mathcal{M}}(\pi^{1:H})}(\widehat{c}_h, \gamma_h) \leftarrow \sum_{p_h, a_h} \mathbb{P}_h^{\widehat{\mathcal{M}}(\pi^{1:H})}(p_h \mid \widehat{c}_h)\gamma_h(a_h \mid p_h)\widehat{r}_{i,h}^{\widehat{\mathcal{M}}(\pi^{1:H})}(\widehat{c}_h, p_h, a_h),$$

where we recall $\gamma_h(a_h \mid p_h) := \prod_{j=1}^n \gamma_{j,h}(a_{j,h} \mid p_{j,h})$.
return $\widehat{\mathcal{M}}(\pi^{1:H}) := (\{\widehat{\mathcal{C}}_h\}_{h\in[H+1]}, \{\widehat{\phi}_{h+1}\}_{h\in[H]}, \{\mathbb{P}_h^{\widehat{\mathcal{M}}(\pi^{1:H}),z}\}_{h\in[H]}, \Gamma, \widehat{r})$

Algorithm 6 $\texttt{Plam}(\pi^{1:H}, \{\widehat{\mathcal{C}}_h\}_{h\in[H+1]}, \{\widehat{\phi}_{h+1}\}_{h\in[H]}, \Gamma, \zeta_1, \zeta_2, \theta_1, \theta_2, \delta_1, \epsilon_e)$: **P**lanning in **l**earned **a**pproximate **m**odel

Input: $\pi^{1:H}, \{\widehat{\mathcal{C}}_h\}_{h\in[H+1]}, \{\widehat{\phi}_{h+1}\}_{h\in[H]}, \Gamma, \zeta_1, \zeta_2, \theta_1, \theta_2, \delta_1, \epsilon_e$
$\widehat{\mathcal{M}}(\pi^{1:H}) \leftarrow \texttt{Construct}(\pi^{1:H}, \{\widehat{\mathcal{C}}_h\}_{h\in[H+1]}, \{\widehat{\phi}_{h+1}\}_{h\in[H]}, \Gamma, \zeta_1, \zeta_2, \theta_1, \theta_2, \delta_1)$ // i.e., Algorithm 5
$\pi^\star \leftarrow \texttt{VIACM}(\widehat{\mathcal{M}}(\pi^{1:H}), \epsilon_e)$ // i.e., Algorithm 3
return $\left\{\pi^\star, \widehat{\mathcal{M}}(\pi^{1:H})\right\}$

- ***State controlled by one controller with asymmetric*** $d = \text{poly}(\log H)$***-step delayed sharing sharing:*** $(AO)^{C(\gamma^{-4}\log\frac{SHO}{\gamma\epsilon} + d)} \log\frac{1}{\delta}$ *for some constant* $C > 0$.

Algorithm 7 $\mathtt{PoS}(\{\widehat{\mathcal{M}}(\pi^{1:H,m})\}_{m\in[K]}, \{\pi^{\star,j}\}_{j\in[K]}, \epsilon_e, N_2)$: **Po**licy **S**election

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

Algorithm 8 $\mathtt{PoS\text{-}Dec}(\{\pi^{\star,j}\}_{j\in[K]}, N_2)$: **Po**licy **S**election for **Dec**-POMDP

Input: $\{\pi^{\star,j}\}_{j\in[K]}, N_2$
for $j \in [K]$ **do**
  Execute $\pi^{\star,j}$ for N_2 trajectories and let the mean accumulated reward be R^j
end for
$\widehat{j} \leftarrow \arg\max_{j\in[K]} R^j$
return $\pi^{\star,\widehat{j}}$

Algorithm 9 $\mathtt{LACI}(\mathcal{G}, \{\widehat{\mathcal{C}}_h\}_{h\in[H+1]}, \{\widehat{\phi}_{h+1}\}_{h\in[H]}, \Gamma, \widehat{L}, \epsilon, \delta_2, \zeta_1, \zeta_2, \theta_1, \theta_2, \delta_1, N_2, \epsilon_e)$: **L**earning with **A**pproximate **C**ommon **I**nformation

Input: $\mathcal{G}, \{\widehat{\mathcal{C}}_h\}_{h\in[H+1]}, \{\widehat{\phi}_{h+1}\}_{h\in[H]}, \Gamma, \widehat{L}, \epsilon, \delta_2, \zeta_1, \zeta_2, \theta_1, \theta_2, \delta_1, N_2, \epsilon_e$
$\{\pi^{1:H,j}\}_{j=1}^K \leftarrow \mathtt{BaSeCAMP}(\mathcal{G}, \widehat{L}, \epsilon, \delta_2)$ // i.e., Algorithm 3 of Golowich et al. (2022a)
for $j \in [K]$ **do**
  $\left\{\pi^{\star,j}, \widehat{\mathcal{M}}(\pi^{1:H,j})\right\} \leftarrow \mathtt{Plam}(\pi^{1:H,j}, \{\widehat{\mathcal{C}}_h\}_{h\in[H+1]}, \{\widehat{\phi}_{h+1}\}_{h\in[H]}, \Gamma, \zeta_1, \zeta_2, \theta_1, \theta_2, \delta_1, \epsilon_e)$ // i.e., Algorithm 6
end for
if learning the equilibrium **then**
  $\pi^{\star,\widehat{j}} \leftarrow \mathtt{PoS}(\{\widehat{\mathcal{M}}(\pi^{1:H,j})\}_{j=1}^K, \{\pi^{\star,j}\}_{j=1}^K, \epsilon_e, N_2)$ // i.e., Algorithm 7
else if learning the team-optimum **then**
  $\pi^{\star,\widehat{j}} \leftarrow \mathtt{PoS\text{-}Dec}(\{\pi^{\star,j}\}_{j=1}^K, N_2)$ // i.e., Algorithm 8
end if
return $\pi^{\star,\widehat{j}}$

- ***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* $C > 0$.

Algorithm 10 ADPNIS($\mathbb{P}_h^{\mathcal{M},c}(\cdot,\cdot|\widehat{c}_h)$): Agent-based **D**ynamic **P**rogramming under **N**ested **I**nformation-**S**haring

Input: $\mathbb{P}_h^{\mathcal{M},c}(\cdot,\cdot|\widehat{c}_h)$

Initialize $V_{n+1}(p_{1:n+1,h}, a_{1:n,h}) \leftarrow \mathbb{E}_{s_h\sim\mathbb{P}_h^{\mathcal{M},c}(\cdot|\widehat{c}_h,p_h), s_{h+1}\sim\mathbb{T}_h(\cdot|s_h,a_h)}\ \mathbb{E}_{o_{h+1}\sim\mathbb{O}_{h+1}(\cdot|s_{h+1})}\Big[r_h(s_h,a_h) + V_{h+1}^{\star,\mathcal{M}}(\widehat{c}_{h+1})\Big]$ for any $p_h \in \mathcal{P}_h$, $a_h \in \mathcal{A}$

for $i = n, \cdots, 1$ **do**

  for each $p_{1:i,h} \in \times_{j=1}^{i}\mathcal{P}_{j,h}$, $a_{1:i-1,h} \in \times_{j=1}^{i-1}\mathcal{A}_j$ **do**

$$u_i^\star(p_{1:i,h}, a_{1:i-1,h}) \leftarrow \arg\max_{a_{i,h}\in\mathcal{A}_i} \mathbb{E}_{p_{i+1,h}\sim\mathbb{P}_h^{\mathcal{M},c}(\cdot|\widehat{c}_h,p_{1:i,h})} V_{i+1}(p_{1:(i+1),h}, a_{1:i,h})$$

$$V_i(p_{1:i,h}, a_{1:i-1,h}) \leftarrow \max_{a_{i,h}\in\mathcal{A}_i} \mathbb{E}_{p_{i+1,h}\sim\mathbb{P}_h^{\mathcal{M},c}(\cdot|\widehat{c}_h,p_{1:i,h})} V_{i+1}(p_{1:(i+1),h}, a_{1:i,h})$$

  end for

end for

for $i = 1, \cdots, n$ **do**

  for each $p_{i,h} \in \mathcal{P}_{i,h}$ **do**

    for each $j = 1, \cdots, i$ **do**

      $p_{j,h} \leftarrow Y_h^{ij}(p_{i,h})$

      $a_{j,h}^\star \leftarrow u_j^\star(p_{1:j,h}, a_{1:j-1}^\star)$

    end for

    $\gamma_{i,h}^\star(p_{i,h}) \leftarrow a_{i,h}^\star$

  end for

end for

return $\{\gamma_{i,h}^\star\}_{i\in[n]}$

- ***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 Missing details in Section 3.1

Before proving Proposition 1, we present some hardness results for solving the stronger solution concepts of team-optimal policy in Dec-POMDPs to further justify the necessity of some favorable information-sharing structures.

Proposition 3. *With 1-step delayed information-sharing structure and Assumption 2, computing the team optimal policy in Dec-POMDPs with $n = 2$ is* NP-hard*.*

To prove Proposition 3, 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 and 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 4. *Without any information sharing, computing jointly team optimal policies in Dec-POMDP with $H = 1$, $n = 2$ is* $\texttt{NP-hard}$.

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. To begin with, we define $c_{\max} = \max_{y_1, y_2, u_1, u_2} c(y_1, y_2, u_1, u_2)$. For any team decision problem, we can construct the following Dec-POMDP:

- $\mathcal{A}_i = \mathcal{U}_i, i = 1, 2$;
- $\mathcal{O}_i = \mathcal{Y}_i, i = 1, 2$;
- $\mathcal{S} = \mathcal{O}_1 \times \mathcal{O}_2$.
- $\mathbb{O}(o_{1,h}, o_{2,h} \mid s_h) = 1$ if $s_h = (o_{1,h}, o_{2,h})$, else 0, for $h \in \{1, 2\}$;
- $r_1(s_1, a_1) = 1 - c(y_1, y_2, u_1, u_2)/c_{\max}$, where $s_1 = (y_1, y_2)$;
- $\mu_1(s_1) = p(y_1, y_2)$, where $s_1 = (y_1, y_2)$.

Based on the construction, computing the optimal policies $\{\pi^\star_{1,1}, \pi^\star_{2,1}\}$ under the no-information-sharing structure in the reduced Dec-POMDP problem will give us the optimal policies $\{\gamma_1^\star, \gamma_2^\star\}$ in the original team decision problem. Concretely, we can construct the optimal policy for the team decision problem as $\gamma_i^\star(y_i) = \pi^\star_{i,1}(o_{i,1})$, where $o_{i,1} = y_i$. Given the $\texttt{NP-hardness}$ of the team decision problem shown in Tsitsiklis and Athans (1985), solving this corresponding Dec-POMDP without information sharing is also $\texttt{NP-hard}$. $\square$

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

Proposition 5. *With 1-step delayed information-sharing structure, computing jointly team optimal policies in Dec-POMDPs with $n = 2$ is at least* $\texttt{NP-hard}$.

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 the proof of Proposition 4, which concerns exactly such a case, computing joint optimal policies in Dec-POMDPs with $n = 2$ is also at least $\texttt{NP-hard}$. $\square$

Finally, we are ready to prove Proposition 3.

Proof of Proposition 3. Similar to the proof of Proposition 5, it suffices to show that the proposition holds for Dec-POMDPs, with $H = 1$ and without information sharing. Note that in the proof of Proposition 4, the constructed Dec-POMDPs has the state space defined as the joint observation space (the Cartesian product of the individual observation spaces), 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 $\gamma = 1$. $\square$

Now let us restate and prove our hardness results regarding NE/CE/CCE in Proposition 1 as the following two propositions.

Proposition 6. *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 POSG with desired information-sharing strategy gives the ϵ-optimal policy in the original POMDP. Given the known PSPACE-hardness of POMDPs (Papadimitriou and Tsitsiklis, 1987; Lusena et al., 2001), we conclude our proof. □

Proposition 7. *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 6, given any instance of a POMDP, 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 $\mathcal{P} = (\mathcal{S}^{\mathcal{P}}, \mathcal{A}^{\mathcal{P}}, \mathcal{O}^{\mathcal{P}}, \{\mathbb{O}_h^{\mathcal{P}}\}_{h\in[H+1]}, \{\mathbb{T}_h^{\mathcal{P}}\}_{h\in[H]}, r^{\mathcal{P}})$, 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

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

- For any $h \in [H]$, $a_{1,h} \in \mathcal{A}_1$, $a_{2,h} \in \mathcal{A}_2$, $s_h, s_{h+1} \in \mathcal{S}$, it holds that $\mathbb{T}_h(s_{h+1} \mid s_h, a_{1,h}, a_{2,h}) = \mathbb{T}_h^{\mathcal{P}}(s_{h+1} \mid s_h, a_{1,h})$;
- For the reward, we use the reward from the original POMDP.

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

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

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 regarding usually how large $C_h P_h$ is.

Proof of Lemma 1. Fix any $h \in [H+1]$. If each player has perfect recall, then it holds that for any joint history $\{o_1, a_1, o_2, \cdots, a_{h-1}, o_h\} \in \mathcal{O}^h \times \mathcal{A}^{h-1}$, there exists some $c_h \in \mathcal{C}_h$ and $p_h \in \mathcal{P}_h$ such that $\{c_h, p_h\} = \{o_1, a_1, o_2, \cdots, a_{h-1}, o_h\}$, which can be found by the functions f_h and g_h introduced after Assumption 1. Therefore, we conclude that $\mathcal{O}^h \times \mathcal{A}^{h-1} \subseteq \mathcal{C}_h \times \mathcal{P}_h$, implying that $C_h P_h \ge (OA)^{h-1}$. □

E.2 Missing details in Section 3.2

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. Specifically, under Assumption 3, we can have the Bellman equation as follows

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

With Assumption 3, we are ready to present our Algorithm 1 based on value iteration in 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.

Implementing the equilibrium subroutine at each step. Now we will discuss the three equilibrium or best response (BR) subroutines at each step $h \in [H]$, 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 computational tractability. To find efficient implementation for these subroutines, we need the following important properties on the prescription-value function.

Proposition 8. *$Q_{i,h}^{\star,\mathcal{G}}(c_h,\gamma_{1,h},\cdots,\gamma_{n,h})$ defined in Algorithm 1 is linear with respect to each $\gamma_{i,h}$. More specifically, we have:*

$$\begin{aligned}\frac{\partial Q_{i,h}^{\star,\mathcal{G}}(c_h,\gamma_{1,h},\cdots,\gamma_{n,h})}{\partial\gamma_{i,h}(a_{i,h}\,|\,p_{i,h})} &= \sum_{s_h',p_{-i,h}'}\sum_{a_{-i,h}'}\mathbb{P}_h^{\mathcal{G}}(s_h',p_{i,h},p_{-i,h}'\,|\,c_h)\gamma_{-i,h}(a_{-i,h}'\,|\,p_{-i,h}')\\ &\quad\times\left(\sum_{o_{h+1},s_{h+1}'}\mathbb{O}_{h+1}(o_{h+1}|s_{h+1}')\mathbb{T}_h(s_{h+1}'|s_h',a_h)\Big[r_{i,h}(s_h,a_h)+V_{i,h+1}^{\star,\mathcal{G}}(c_{h+1})\Big]\right).\end{aligned} \tag{E.1}$$

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 that $\gamma_{i,h}$ 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:

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

  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 $A_iP_{i,h}$ 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 $\texttt{poly}(S,A,P_h,\frac{1}{\epsilon_e})$ 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, denoted as the NE-BR subroutine, is defined as follows: it outputs 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, a uniform mixture of $\{\gamma_{1,h}^{\star,t}, \cdots, \gamma_{n,h}^{\star,t}\}_{t=1}^{T}$ up to some error ϵ_e, which satisfy for any $i \in [n]$:

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

  This subroutine can be implemented using standard no-external-regret learning algorithm as in Gordon et al. (2008); Farina et al. (2022) with $\texttt{poly}(S, A, P_h, \frac{1}{\epsilon_e})$ time.

  Similarly, the CCE-BR subroutine can be defined as follows: it outputs the best response $\gamma_{i,h}^{\star}$ of agent i, 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 for any $i \in [n]$:

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

  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 $(u_{i,h} \diamond \gamma_{i,h})(a_{i,h} \mid p_{i,h}) := \sum_{u_{i,h,p_{i,h}}(a'_{i,h})=a_{i,h}} \gamma_{i,h}(a'_{i,h} \mid p_{i,h})$. 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 the function $Q_{i,h}^{\star}(c_h, \gamma_{1,h}, \cdots, \gamma_{n,h})$ is concave (in fact, linear) with respect to each $\gamma_{i,h}$, one can run the *no-linear-regret* algorithm as in Gordon et al. (2008), such that the time-averaged policy will give us the approximate CE. In particular, such guarantee can be achieved by running swap regret minimization algorithm in Blum and Mansour (2007) separately for each $(i, p_{i,h})$ and the corresponding time complexity will be of $\texttt{poly}(S, A, P_h, \frac{1}{\epsilon_e})$.

  The CE-BR subroutine can be defined as follows: it will output the best strategy modification $u_{i,h}^{\star}$ of agent i, 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 from that of CE except that only the agent i runs the *no-linear-regret* algorithm.

E.3 Proof of Theorem 2

To prove Theorem 1, we prove our main theorem, Theorem 2, which is a generalized version. We will first bound the sub-optimality of the planning algorithm on $\mathcal{M}$ at each step h through the following two lemmas.

Lemma 6. *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

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

where Equation (E.2) comes from the inductive hypothesis, Equation (E.3) holds since $\widehat{\pi}_h^{\star}(\cdot\mid\cdot,\widehat{c}_h,\cdot)$ is an ϵ_e-NE/CCE for the stage game and $V_{i,h+1}^{\widehat{\pi}^{\star},\mathcal{M}}(c_{h+1}) = V_{i,h+1}^{\star,\mathcal{M}}(\widehat{c}_{h+1})$ through a simple induction argument using Definition 13. ☐

Corollary 1. *Fix the input $\mathcal{M}$, $i \in [n]$, $\widehat{\pi}$ whose $\widehat{\pi}_{j,h} : \Omega_h \times \mathcal{P}_{j,h} \times \widehat{\mathcal{C}}_h \to \Delta(\mathcal{A}_j)$ for $j \in [n]$ takes only approximate common information instead of the exact common information as the input, and $\epsilon_e > 0$ for Algorithm 4. For any $h \in [H+1]$, $c_h \in \mathcal{C}_h$, and $\pi_i \in \Pi_i$, the output of Algorithm 4, $\widehat{\pi}_i^{\star}$ satisfies that*

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

Proof. Since Algorithm 4 simply replaces the equilibrium oracle at each stage of Algorithm 3 by a best-response oracle, its proof follows directly from the proof of Lemma 6. ☐

Lemma 7. *For any $h \in [H+1]$, $c_h \in \mathcal{C}_h$, 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. It is direct to see the lemma holds for the step $H+1$. For step $h \in [H]$, it holds that

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

where 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 $\widehat{\pi}_h^\star(\cdot\,|\,\cdot, \widehat{c}_h, \cdot)$ is an ϵ_e-CE for the stage game. □

Corollary 2. *Fix the input $\mathcal{M}$, $i \in [n]$, $\widehat{\pi}$ whose $\widehat{\pi}_{j,h} : \Omega_h \times \mathcal{P}_{j,h} \times \widehat{\mathcal{C}}_h \to \Delta(\mathcal{A}_j)$ for $j \in [n]$ takes only approximate common information instead of the exact common information as the input, and $\epsilon_e > 0$ for Algorithm 4. For any $h \in [H+1]$, $c_h \in \mathcal{C}_h$, and $\phi_i \in \Phi_i$, the output of Algorithm 4, $\widehat{\pi}_i^\star$ satisfies that*

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

Proof. Similar to the proof of Corollary 1, its proof follows directly from Lemma 7. □

Now we prove Lemma 2, showing the difference between the approximate value functions and true value functions under the same set of policies.

Proof of Lemma 2. Note that it suffices to consider any policy $\pi' \in \Pi^{\mathrm{det}}$ instead of $\pi' \in \Delta(\Pi^{\mathrm{det}})$. Obviously, the proposition holds for $h = H+1$. For step $h \in [H]$, we have

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

where $\bar{\pi} \in \Delta(\Pi^{\mathrm{det}})$ is the policy following π' from step 1 to $h-1$ and π from step h to H, thus completing the proof. □

Finally, we are ready to prove our main theorem, Theorem 2. Before that, we need to show that the equilibrium subroutine at each $h \in [H]$ in $\mathcal{M}$ is also computationally tractable. Specifically, similar to Proposition 8, we can show $Q_{i,h}^{\star, \mathcal{M}}$ is also linear w.r.t. each $\gamma_{i,h}$ for $i \in [n]$. Hence, the algorithms developed to implement the equilibrium subroutine for $\mathcal{G}$ in Appendix E.2 are directly applicable for $\mathcal{M}$.

Proposition 9. *Given $\mathcal{M}$ that is consistent with the approximate belief $\{\mathbb{P}_h^{\mathcal{M},c}(s_h, p_h \mid \widehat{c}_h)\}_{h\in[H]}$, we have $Q_{i,h}^{\star,\mathcal{M}}(\widehat{c}_h, \gamma_{1,h}, \cdots, \gamma_{n,h})$ defined in Algorithm 3 is linear with respect to each $\gamma_{i,h}$. More specifically, we have:*

$$\frac{\partial Q_{i,h}^{\star,\mathcal{M}}(\widehat{c}_h, \gamma_{1,h}, \cdots, \gamma_{n,h})}{\partial \gamma_{i,h}(a_{i,h} \mid p_{i,h})} = \sum_{s_h', p_{-i,h}'} \sum_{a_{-i,h}'} \mathbb{P}_h^{\mathcal{M},c}(s_h', p_{i,h}, p_{-i,h}' \mid \widehat{c}_h) \gamma_{-i,h}(a_{-i,h}' \mid p_{-i,h}') \times \left(\sum_{o_{h+1}, s_{h+1}'} \mathbb{O}_{h+1}(o_{h+1} \mid s_{h+1}') \mathbb{T}_h(s_{h+1}' \mid s_h', a_h) \left[r_{i,h}(s_h, a_h) + V_{i,h+1}^{\star,\mathcal{M}}(\widehat{c}_{h+1}) \right] \right). \tag{E.6}$$

Proof. The partial derivative can be easily verified by algebraic manipulations and the definition of $Q_{i,h}^{\star,\mathcal{M}}$. From Equation (E.6), we could notice that $\gamma_{i,h}$ does not appear on the RHS, which proves $Q_{i,h}^{\star,\mathcal{M}}(\widehat{c}_h, \gamma_{1,h}, \cdots, \gamma_{n,h})$ is linear with respect to $\gamma_{i,h}$. □

Proof of Theorem 2. For computing NE/CCE, we define for each agent $i \in [n]$

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

Now note that

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

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

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

Now note that

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

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 computational complexity. A major difference from the exact common-information setting is that it is unclear whether there exist efficient NE/CE/CCE subroutines at each step h. However, if $\mathcal{M}$ is consistent with some approximate belief $\{\mathbb{P}_h^{\mathcal{M},c}(s_h, p_h \mid \widehat{c}_h)\}_{h\in[H]}$, by Proposition 9, we conclude the NE subroutine for zero-sum or cooperative games and CE/CCE subroutine for general-sum games can be also implemented efficiently with the computational complexity of $\texttt{poly}(S, A, P_h, \frac{1}{\epsilon_e})$. Hence the overall computational complexity of the Algorithm 3 is $H \max_h \widehat{C}_h \texttt{poly}(S, A, P_h, \frac{1}{\epsilon_e})$, where $\widehat{C}_h$ comes from the loop at each step h. □

Finally, we are ready to prove Theorem 1 as a special case.

Proof of Theorem 1. we can leverage the reduction in Nayyar et al. (2013a) that reduces $\mathcal{G}$ to an *exact* common information model $\mathcal{M}(\mathcal{G})$ 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 $h \in [H+1], c_h \in \mathcal{C}_h$, and $\mathcal{M}(\mathcal{G})$ is consistent with $\{\mathbb{P}_h^{\mathcal{G}}(s_h, p_h \mid c_h)\}_{h\in[H]}$. Therefore, by applying Theorem 2, we conclude the proof. □

E.4 Proof of Theorem 3

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 7. 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 *finite-memory truncation* to compress the common information, and indeed, the corresponding most recent L steps of history is a kind of approximate common information. Importantly, we need the a series of following result showing that the most recent history is enough to predict the latent state of the POSG (with information sharing).

Lemma 8 (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*

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

where we recall the definition of $\mathbf{B}_h$ in Appendix A.1.

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

Lemma 9. *Given two finite domains X, Y, and the conditional probability $q(y \mid x)$ 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*

$$F^q(P; y)(x) = \frac{P(x)q(y \mid x)}{\sum_{x'\in X} P(x')q(y \mid x')}. \tag{E.7}$$

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

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

Proof. This is a direct consequence of 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 3. *Suppose $b, b' \in \Delta(\mathcal{S})$ with $b \ll b'$, and fix any $h \in [H], i \in [n]$. Then*

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

Lemma 10 (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 $\pi' \in \Delta(\Pi^{\mathrm{det}})$ 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*

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

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

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

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 9, it holds that

$$\mathbb{E}^{\mathcal{G}}_{\pi'}\mathbb{E}_{y\sim q\cdot\boldsymbol{b}_h(a_{1:h-1},o_{1:h})}\|F^q(\boldsymbol{b}_h(a_{1:h-1},o_{1:h});y) - F^q(\boldsymbol{b}'_h(a_{h-L:h-1},o_{h-L+1:h});y)\|_1 \le \epsilon. \tag{E.11}$$

Proof. Equation (E.8) is from Theorem 4.7 in Golowich et al. (2022b). For the remaining, it suffices to only consider $\pi' \in \Pi^{\mathrm{det}}$. We prove Equation (E.9) first. Note that if $h-L \le 1$, then we have $\boldsymbol{b}_h(a_{1:h-1},o_{1:h-1}) = \boldsymbol{b}'_h(a_{h-L:h-1},o_{h-L+1:h-1})$. The proposition holds trivially. Now let us consider $h > L+1$. Fix some history $(a_{1:h-L-1},o_{1:h-L})$. We condition on this history throughout the proof. For $0 \le t \le L$, define the random variables

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

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 $0 \leq t \leq L-1$, by denoting the shorthand notation of the matrix $A := \mathbb{T}_{h-L+t}(a_{h-L+t})$, we have :

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

where the second last step comes from the data processing inequality and the last step comes from Lemma 8. By induction and the choice of L, we have that

$$
\mathbb{E}_{o_{h-L:h-1}, a_{h-L:h-1} \sim \pi'} \sqrt{\exp\left(\frac{D_2\left(b_h \| b'_h\right)}{4}\right) - 1} \leq \left(1 - \frac{\gamma^4}{2^{40}}\right)^L S \leq \frac{\epsilon}{4}. \tag{E.12}
$$

It follows from Lemma 10 that

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

Equation (E.9) follows from Equation (E.12) and Lemma 8. Equation (E.10) follows from Equation (E.12) and Corollary 3. Equation (E.11) follows from Equation (E.12) and Lemma 9. □

Before instantiating the information structure in particular cases, we prove Lemma 3 first, which is a more sufficient condition for our Definition 7.

Proof of Lemma 3. By Definition 8, it holds that

$$
\left|\mathbb{E}^{\mathcal{G}}[r_{i,h}(s_h, a_h) \mid c_h, \gamma_h] - \widehat{r}_{i,h}^{\mathcal{M}}(\widehat{c}_h, \gamma_h)\right| \leq \sum_{s_h, a_h} \left|\mathbb{P}_h^{\mathcal{G}}(s_h, a_h \mid c_h, \gamma_h) - \mathbb{P}_h^{\mathcal{M}, o}(s_h, a_h \mid \widehat{c}_h, \gamma_h)\right|.
$$

Therefore, it suffices to bound the right hand size to order to prove Equation (6.2). Now, note that for any $c_h \in \mathcal{C}_h, \gamma_h \in \Gamma_h$:

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

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

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

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 $\widetilde{r}_{i,h}^{\mathcal{M}}(\widehat{c}_h,\gamma_h)$ will follow from the consistency condition (4.4) and (4.5). Now we will show when $\mathcal{G}$ satisfies our Assumptions 1, 2, 3, how we can construct approximate common information with history truncation that satisfies Definition 7.

One-step delayed information-sharing. In this case, the information structure has $c_h = \{a_{1:h-1},o_{1:h-1}\}$, $p_{i,h}=\{o_{i,h}\}$, $z_{h+1}=\{o_h,a_h\}$, and $\mathbb{P}_h^{\mathcal{G}}(s_h,p_h\,|\,c_h)=\boldsymbol{b}_h(a_{1:h-1},o_{1:h-1})(s_h)\mathbb{O}_h(o_h\,|\,s_h)$, which verifies Assumption 3. 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 $\mathbb{P}_h^{\mathcal{M},c}(s_h,p_h\,|\,\widehat{c}_h)=\boldsymbol{b}'_h(a_{h-L:h-1},o_{h-L+1:h-1})(s_h)\mathbb{O}_h(o_h\,|\,s_h)$. Now we are ready to verify that it satisfies Definition 7.

- Obviously, it satisfies condition (4.1).
- Note that for any $c_h\in\mathcal{C}_h$ and the corresponding $\widehat{c}_h$ constructed above:

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

  Therefore, by setting $L\geq C\gamma^{-4}\log(\frac{S}{\epsilon})$, according to Equation (E.8) in Theorem 10, we conclude that for any $\pi'\in\Pi^{\mathrm{det}},h\in[H]$:

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

  Therefore, conditions (4.2), (4.3) in Definition 7 are satisfied using Lemma 3 with $\epsilon_r=\epsilon_z=\epsilon$.

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 computes an ϵ-NE if the POSG is zero-sum or cooperative, and an ϵ-CE/CCE if the POSG is general-sum with time complexity $H(AO)^{C\gamma^{-4}\log\frac{SH}{\epsilon}}poly(S,A,O,H,\frac{1}{\epsilon})$ for some universal constant $C>0$.*

Proof. It is direct to see that $\widehat{C}_h\leq(AO)^L$ and $P_h\leq O$, 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 computation subroutines. □

State controlled by one controller with asymmetric delay sharing. The information structure is given as $c_h = \{o_{1,1:h}, o_{2,1:h-d}, a_{1,1:h-1}\}$, $p_{1,h} = \emptyset$, $p_{2,h} = \{o_{2,h-d+1:h}\}$, $z_{h+1} = \{o_{1,h+1}, o_{2,h-d+1}, a_h\}$. It is a bit less straightforward to verify Assumption 3. We do so by explicitly computing $\mathbb{P}_h^{\mathcal{G}}(s_h, p_h \mid c_h)$ as follows. Denote $\tau_{h-d} = \{a_{1:h-d-1}, o_{1:h-d}\}$, $f_a = \{a_{1,h-d:h-1}\}$, $f_o = \{o_{1,h-d+1:h}\}$. Now $\mathbb{P}_h^{\mathcal{G}}(s_h, p_h \mid c_h) = \sum_{s_{h-d}} \mathbb{P}_h^{\mathcal{G}}(s_h, p_h \mid s_{h-d}, f_a, f_o)\mathbb{P}_h^{\mathcal{G}}(s_{h-d} \mid \tau_{h-d}, f_a, f_o)$. It is direct to see that $\mathbb{P}_h^{\mathcal{G}}(s_h, p_h \mid s_{h-d}, f_a, f_o)$ does not depend on the policy. For $\mathbb{P}_h^{\mathcal{G}}(s_{h-d} \mid \tau_{h-d}, f_a, f_o)$, the following holds

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

Now note that

$$\begin{aligned}
&\mathbb{P}_h^{\mathcal{G}}(s_{h-d}, f_a, f_o \mid \tau_{h-d}) \\
&\quad = \boldsymbol{b}_{h-d}(a_{1:h-d-1}, o_{1:h-d})(s_{h-d})\mathbb{P}_h^{\mathcal{G}}(a_{1,h-d} \mid \tau_{h-d})\mathbb{P}_h^{\mathcal{G}}(o_{1,h-d+1} \mid s_{h-d}, a_{1,h-d}) \cdots \mathbb{P}_h^{\mathcal{G}}(o_{1,h} \mid s_{h-d}, a_{1,h-d:h-1}).
\end{aligned}$$

Now let us use the notation $P_h(f_o \mid s_{h-d}, f_a) := \prod_{t=1}^{d} \mathbb{P}_h^{\mathcal{G}}(o_{1,h-d+t} \mid s_{h-d}, a_{1,h-d:h-d+t-1})$. Then it holds that $\sum_{f_o} P_h(f_o \mid s_{h-d}, f_a) = 1$, which suggests that the notation $P_h(f_o \mid s_{s-d}, f_a)$ can be understood as a conditional probability. With such notation, we have

$$\begin{aligned}
\mathbb{P}_h^{\mathcal{G}}(s_{h-d} \mid \tau_{h-d}, f_a, f_o) &= \frac{\boldsymbol{b}_{h-d}(a_{1:h-d-1}, o_{1:h-d})(s_{h-d})P_h(f_o \mid s_{h-d}, f_a)}{\sum_{s'_{h-d}} \boldsymbol{b}_{h-d}(a_{1:h-d-1}, o_{1:h-d})(s'_{h-d})P_h(f_o \mid s'_{h-d}, f_a)} \\
&= F^{P_h(\cdot \mid \cdot, f_a)}(\boldsymbol{b}_{h-d}(a_{1:h-d-1}, o_{1:h-d}); f_o)(s_{h-d}),
\end{aligned}$$

where we recall the definition of F in Lemma 9. Finally, we compute:

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

It is easy to see this expression does not depend on the policy executed, thus verifying Assumption 3. 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:

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

To verify Definition 7:

- Obviously, it satisfies the condition (4.1).
- For any $c_h \in \mathcal{C}_h$ and the corresponding $\widehat{c}_h$ constructed above:

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

Finally, for any policy $\pi' \in \Pi^{\mathrm{det}}$ taking expectations over τ_{h-d}, f_a, f_o, we conclude that as long as $L \geq C\gamma^{-4} \log \frac{S}{\epsilon}$ using Equation E.11 of Theorem 10, we have

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

Therefore, conditions (4.2), (4.3) in Definition 7 are satisfied using Lemma 3 with $\epsilon_r = \epsilon_z = \epsilon$.

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 computes an ϵ-NE if the POSG is zero-sum or cooperative, and an ϵ-CE/CCE if the POSG is general-sum with 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 $\mathbb{P}_h^{\mathcal{M},c}(s_h, p_h \mid \widehat{c}_h)$ and the equilibrium computation subroutines. $\square$

Information sharing with one-directional-one-step delay. For this case, we have $c_h = \{a_{1:h-1}, o_{1:h-1}, o_{1,h}\}$, $p_{1,h} = \emptyset$, $p_{2,h} = \{o_{2,h}\}$, $z_{h+1} = \{o_{1,h+1}, o_{2,h}, a_h\}$, and $\mathbb{P}_h^{\mathcal{G}}(s_h, p_h \mid c_h) = \boldsymbol{b}_h(a_{1:h-1}, o_{1:h-1}, o_{1,h})(s_h)\mathbb{P}_h(o_{2,h} \mid s_h, o_{1,h})$, where $\mathbb{P}_h(o_{2,h} \mid s_h, o_{1,h}) = \frac{\mathbb{O}_h(o_{1,h}, o_{2,h} \mid s_h)}{\sum_{o_{2,h}'}\mathbb{O}_h(o_{1,h}, o_{2,h}' \mid s_h)}$, thus verifying Assumption 3. 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, we define the belief as

$$\mathbb{P}_h^{\mathcal{M},c}(s_h, p_h \mid \widehat{c}_h) = \boldsymbol{b}_h'(a_{h-L:h-1}, o_{h-L+1:h-1}, o_{1,h})(s_h)\mathbb{P}_h(o_{2,h} \mid s_h, o_{1,h}).$$

Now we are ready to verify that Definition 7 is satisfied.

- Obviously, the condition (4.1) is satisfied.
- Note that for any $c_h \in \mathcal{C}_h$ and the corresponding $\widehat{c}_h$ constructed above:

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

  Therefore, by setting $L \ge C\gamma^{-4}\log(\frac{S}{\epsilon})$, according to (E.10) in Theorem 10, we conclude that for any $\pi' \in \Pi^{\mathrm{det}}$:

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

  Therefore, conditions (4.2), (4.3) in Definition 7 are satisfied using Lemma 3 with $\epsilon_r = \epsilon_z = \epsilon$.

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 computes an ϵ-NE if the POSG is zero-sum or cooperative, and an ϵ-CE/CCE if the POSG is general-sum with 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 direct to see that $\widehat{C}_h \le (AO)^L$ and $P_h \le O_2$. The polynomial dependence on S, H, A, and O comes from computing $\mathbb{P}_h^{\mathcal{M},c}(s_h, p_h \mid \widehat{c}_h)$ and the equilibrium computation subroutines. $\square$

Uncontrolled state process with delayed sharing. As long as the state transition does not depend on the actions, Assumption 3 is satisfied. To be more concrete, we have

$$\mathbb{P}_h^{\mathcal{G}}(s_h, p_h \mid c_h) = \sum_{s_{h-d}} \boldsymbol{b}_{h-d}(o_{h-d-L+1:h-d})(s_{h-d})\mathbb{P}_h^{\mathcal{G}}(s_h, o_{h-d+1:h} \mid s_{h-d}),$$

which verifies Assumption 3, where in the notation for $\boldsymbol{b}_{h-d}$, we omit the actions since they do not affect transitions. For generality, we consider the d-step delayed sharing information structure, where $d \geq 0$ and not necessarily $d = 1$, as in the one-step delayed information sharing structure. The information structure satisfies $c_h = \{o_{1:h-d}\}$, $p_{i,h} = \{o_{i,h-d+1:h}\}$, and $z_{h+1} = \{o_{h-d+1}\}$. Fix a $L > 0$, the approximate common information is $\widehat{c}_h = \{o_{h-d-L+1:h-d}\}$, the corresponding belief is $\mathbb{P}_h^{\mathcal{M},c}(s_h, p_h \mid \widehat{c}_h) = \sum_{s_{h-d}} \boldsymbol{b}'_{h-d}(o_{h-d-L+1:h-d})(s_{h-d})\mathbb{P}_h^{\mathcal{G}}(s_h, o_{h-d+1:h} \mid s_{h-d})$. Now we are ready to verify Definition 7.

- Obviously, the condition (4.1) is satisfied.
- Note that for any c_h and the corresponding $\widehat{c}_h$ constructed above:

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

  where for the last step, we use Lemma 11 (proved later). Therefore, by setting $L \geq C\gamma^{-4}\log(\frac{S}{\epsilon})$, according to Equation (E.9) in Theorem 10, we conclude that for any $\pi' \in \Pi^{\det}$:

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

  This verifies the conditions (4.2), (4.3) in Definition 7 using Lemma 3 with $\epsilon_r = \epsilon_z = \epsilon$.

Finally, to guarantee that $\widehat{\pi}^\star$ 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 14. *Let $\epsilon, \gamma > 0$. Algorithm 1 given a γ-observable POSG of uncontrolled state process computes an ϵ-NE if the POSG is zero-sum or cooperative, and an ϵ-CE/CCE if the POSG is general-sum with 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 direct to see that $\widehat{C}_h \leq O^L$ and $P_h = O^d$. The polynomial dependence on S, A, H, and O^d comes from computing $\mathbb{P}_h^{\mathcal{M},c}(s_h, p_h \mid \widehat{c}_h)$ and the equilibrium computation subroutines. □

Symmetric information game. For symmetric information games, it has the following information structure: $c_h = \{a_{1:h-1}, o_{1:h}\}$, $p_{i,h} = \emptyset$, $z_{h+1} = \{a_h, o_{h+1}\}$, and $\mathbb{P}_h^{\mathcal{G}}(s_h, p_h \mid c_h) = \boldsymbol{b}_h(a_{1:h-1}, o_{1:h})(s_h)$, verifying Assumption 3. Fix $L > 0$, 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 $\mathbb{P}_h^{\mathcal{M},c}(s_h, p_h \mid \widehat{c}_h) = \boldsymbol{b}'_h(a_{h-L:h-1}, o_{h-L+1:h})(s_h)$. Now we are ready to verify Definition 7.

- Obviously, it satisfies the condition (4.1).

- Note that for any $c_h \in \mathcal{C}_h$ and the corresponding $\widehat{c_h}$ constructed above:

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

  Therefore, by setting $L \geq C\gamma^{-4}\log(\frac{S}{\epsilon})$, according to (E.9) in Theorem 10, we conclude that for any $\pi' \in \Pi^{\text{det}}$:

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

  Therefore, the conditions (4.2) and (4.3) in Definition 7 are satisfied with $\epsilon_r = \epsilon_z = \epsilon$ using Lemma 3.

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 15. *Let $\epsilon, \gamma > 0$. Algorithm 1 given a γ-observable POSG of symmetric information computes an ϵ-NE if the POSG is zero-sum or cooperative, and an ϵ-CE/CCE if the POSG is general-sum with time complexity $H(AO)^{C\gamma^{-4}\log\frac{SH}{\epsilon}}\,poly(S, A, H, O, \frac{1}{\epsilon})$ for some universal constant $C > 0$.*

Proof. It is direct to see that $\widehat{C}_h = (AO)^L$ and $P_h = 1$, 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 computation subroutines. □

We conclude the section by proving the following lemma.

Lemma 11. *For any given sequence $\{x_i\}_{i=1}^m$ and $\{\{y_{i,j}\}_{i=1}^m\}_{j=1}^n$ such that $\sum_{j=1}^n |y_{i,j}| = 1$, $\forall i \in [m]$. The following holds*

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

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

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

Note that $\|\boldsymbol{Y}^\top\|_1 = \|\boldsymbol{Y}\|_\infty = \max_i \sum_{j=1}^n |y_{i,j}| = 1$. Therefore, we have $\sum_{j=1}^n |\sum_{i=1}^m x_i y_{i,j}| \leq \sum_{i=1}^m |x_i|$, and conclude the proof. □

E.5 Proof of Theorem 4

Note that our previous planning algorithms require the knowledge of 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 potentially treat $\mathcal{G}$ as a (fully-observable) Markov game on the state space of c_h, and use black-box algorithms for learning Markov games. However, this formulation could be neither computationally nor sample efficient because of the typical large space of common information. Therefore, we have to learn NE/CE/CCE in the approximate model $\mathcal{M}$ with the state space of $\widehat{c_h}$ in Definition 7. However, the key problem is that we can only sample according to the model of $\mathcal{G}$ instead of $\mathcal{M}$. As we highlighted in Section 4.2 of our main paper, to circumvent this issue, inspired by the idea of Golowich et al. (2022a), one solution is to construct $\widetilde{\mathcal{M}}(\pi^{1:H})$ using a sequence of H policies $\pi^{1:H}$ according to Definition 9, where each $\pi^h \in \Delta(\Pi^{\text{det}})$. Formally, Proposition 10 verifies that $\widetilde{\mathcal{M}}(\pi^{1:H})$ constructed according to Definition 9 can be simulated by executing policies π^h at each step h in the underlying true model $\mathcal{G}$.

Proposition 10. *Given $\widetilde{\mathcal{M}}(\pi^{1:H})$ as in Definition 9, it holds that for any $i \in [n]$, $h \in [H]$, $\widehat{c}_h \in \widehat{\mathcal{C}}_h$, $\gamma_h \in \Gamma_h$, $o_{h+1} \in \mathcal{O}$, $z_{h+1} \in \mathcal{Z}_{h+1}$:*

$$\mathbb{P}_h^{\widetilde{\mathcal{M}}(\pi^{1:H}),z}(z_{h+1} \mid \widehat{c}_h, \gamma_h) = \mathbb{P}_h^{\pi^h_{1:h-1},\mathcal{G}}(z_{h+1} \mid \widehat{c}_h, \gamma_h),$$
$$\widehat{r}_{i,h}^{\widetilde{\mathcal{M}}(\pi^{1:H})}(\widehat{c}_h, \gamma_h) = \mathbb{E}^{\mathcal{G}}_{\pi^h_{1:h-1}}[r_{i,h}(s_h, a_h) \mid \widehat{c}_h, \gamma_h].$$

Proof. Note for $\mathbb{P}_h^{\pi^h_{1:h-1},\mathcal{G}}(z_{h+1} \mid \widehat{c}_h, \gamma_h)$, it holds that

$$\begin{aligned}
&\mathbb{P}_h^{\pi^h_{1:h-1},\mathcal{G}}(z_{h+1} \mid \widehat{c}_h, \gamma_h) \\
&= \sum_{\substack{p_h, a_h, o_{h+1}: \\ \chi_{h+1}(p_h, a_h, o_{h+1}) = z_{h+1}}} \mathbb{P}_h^{\pi^h_{1:h-1},\mathcal{G}}(p_h, a_h, o_{h+1} \mid \widehat{c}_h, \gamma_h) \\
&= \sum_{\substack{s_h, p_h, a_h, o_{h+1}: \\ \chi_{h+1}(p_h, a_h, o_{h+1}) = z_{h+1}}} \left(\mathbb{P}_h^{\pi^h_{1:h-1},\mathcal{G}}(s_h, p_h, \mid \widehat{c}_h)\gamma_h(a_h \mid p_h) \times \sum_{s_{h+1}} \mathbb{T}_h(s_{h+1} \mid s_h, a_h)\mathbb{O}_{h+1}(o_{h+1} \mid s_{h+1})\right),
\end{aligned}$$

where we recall the shorthand notation $\gamma_h(a_h \mid p_h) := \prod_{j \in [n]} \gamma_{j,h}(a_{j,h} \mid p_{j,h})$. Now by Definition 9, we have $\mathbb{P}_h^{\pi^h_{1:h-1},\mathcal{G}}(s_h, p_h, \mid \widehat{c}_h) = \mathbb{P}_h^{\widetilde{\mathcal{M}}(\pi^{1:H}),c}(s_h, p_h, \mid \widehat{c}_h)$. Combined with Equation (4.4) of Definition 8, we conclude $\mathbb{P}_h^{\widetilde{\mathcal{M}}(\pi^{1:H}),z}(z_{h+1} \mid \widehat{c}_h, \gamma_h) = \mathbb{P}_h^{\pi^h_{1:h-1},\mathcal{G}}(z_{h+1} \mid \widehat{c}_h, \gamma_h)$. At the same time, we can prove $\widehat{r}_{i,h}^{\widetilde{\mathcal{M}}(\pi^{1:H})}(\widehat{c}_h, \gamma_h) = \mathbb{E}^{\mathcal{G}}_{\pi^h_{1:h-1}}[r_{i,h}(s_h, a_h) \mid \widehat{c}_h, \gamma_h]$ holds by the same derivation. □

Therefore, different from a generic $\mathcal{M}$ in Definition 7, 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* $\{\widehat{\mathcal{C}}_h\}_{h \in [H+1]}$. 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; Golowich et al., 2022a). 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. *Fix $\widehat{L} > 0$ as given in Definition 10. For each $h \in [H]$, $\zeta > 0$, and a joint policy $\pi \in \Delta(\Pi^{det})$, define the set $\mathcal{C}^{low}_{h,\zeta}(\pi) \subseteq \widehat{\mathcal{C}}_h$ as*

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

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

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

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

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

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

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, for any $\phi > 0$, define the under-explored states under some policy $\pi \in \Delta(\Pi^{\det})$ 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 12. *Fix $\widehat{L} > 0$ as given in Definition 10. Fix any $\zeta > 0, \phi > 0, h \in [H]$. Consider any policies π, $\pi' \in \Delta(\Pi^{det})$, 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, we have*

$$d_{\mathcal{C},h}^{\pi,\mathcal{G}}(\mathcal{C}_{h,\zeta}^{low}(\pi')) \leq \frac{A^{2\widehat{L}}O^{\widehat{L}}\zeta}{\phi} + \mathbf{1}[h > \widehat{L}] \cdot d_{\mathcal{S},h-\widehat{L}}^{\pi,\mathcal{G}}(\mathcal{U}_{\phi,h-\widehat{L}}^{\mathcal{G}}(\pi')).$$

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

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

where we recall the definition of $\widehat{f}_h$ and x_h from Definition 10. Therefore, we have

$$\begin{aligned}
\sum_{\widehat{c}_h \notin \mathcal{C}_{h,\zeta}^{\text{low}}(\pi')} d_{\mathcal{C},h}^{\pi,\mathcal{G}}(\widehat{c}_h) &= \sum_{\substack{\widehat{c}_h \notin \mathcal{C}_{h,\zeta}^{\text{low}}(\pi') \\ x_h: \widehat{f}_h(x_h)=\widehat{c}_h}} d_{\mathcal{X},h}^{\pi,\mathcal{G}}(x_h) \\
&= \sum_{x_h: \widehat{f}_h(x_h) \notin \mathcal{C}_{h,\zeta}^{\text{low}}(\pi')} d_{\mathcal{X},h}^{\pi,\mathcal{G}}(x_h) \\
&\geq \sum_{x_h \notin \mathcal{X}_{h,\zeta}^{\text{low}}(\pi')} d_{\mathcal{X},h}^{\pi,\mathcal{G}}(x_h),
\end{aligned}$$

where the last step comes from the fact that $x_h \notin \mathcal{X}_{h,\zeta}^{\text{low}}(\pi')$ implies $\widehat{f}_h(x_h) \notin \mathcal{C}_{h,\zeta}^{\text{low}}(\pi')$. This leads to that

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

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

The next step is to learn $\mathbb{P}_h^{\widetilde{\mathcal{M}}(\pi^{1:H}),z}(z_{h+1} \mid \widehat{c}_h, \gamma_h)$, $\widetilde{r}_{i,h}^{\widetilde{\mathcal{M}}(\pi^{1:H})}(\widehat{c}_h, \gamma_h)$ of the model $\widetilde{\mathcal{M}}(\pi^{1:H})$, which are defined as $\mathbb{P}_h^{\pi_{1:h-1}^h,\mathcal{G}}(z_{h+1} \mid \widehat{c}_h, \gamma_h)$ and $\mathbb{E}_{\pi_{1:h-1}^h}^{\mathcal{G}}[r_{i,h}(s_h, a_h) \mid \widehat{c}_h, \gamma_h]$, respectively. The challenge here compared with the single-agent learning problem (Golowich et al., 2022a) 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 in general lies in continuous spaces, and even if we only consider *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 $\mathbb{P}_h^{\widetilde{\mathcal{M}}(\pi^{1:H}),z}(z_{h+1} \mid \widehat{c}_h, \gamma_h)$ by enumerating all possible $\widehat{c}_h$ and γ_h is not statistically efficient. To circumvent this issue, we observe the fact that for $\mathbb{P}_h^{\widetilde{\mathcal{M}}(\pi^{1:H}),z}(z_{h+1} \mid \widehat{c}_h, \gamma_h)$, it holds that

$$\mathbb{P}_h^{\pi_{1:h-1}^h,\mathcal{G}}(z_{h+1} \mid \widehat{c}_h, \gamma_h) = \sum_{\substack{p_h, a_h, o_{h+1}: \\ \chi_{h+1}(p_h, a_h, o_{h+1}) = z_{h+1}}} \mathbb{P}_h^{\pi_{1:h-1}^h,\mathcal{G}}(p_h, a_h, o_{h+1} \mid \widehat{c}_h, \gamma_h),$$

where we recall χ_{h+1} in Assumption 1. Further, notice the decomposition for $\mathbb{P}_h^{\pi_{1:h-1}^h,\mathcal{G}}(p_h, a_h, o_{h+1} \mid \widehat{c}_h, \gamma_h)$:

$$\mathbb{P}_h^{\pi_{1:h-1}^h,\mathcal{G}}(p_h, a_h, o_{h+1} \mid \widehat{c}_h, \gamma_h) = \mathbb{P}_h^{\pi_{1:h-1}^h,\mathcal{G}}(p_h \mid \widehat{c}_h) \prod_{i=1}^n \gamma_{i,h}(a_{i,h} \mid p_{i,h}) \mathbb{P}_h^{\pi_{1:h-1}^h,\mathcal{G}}(o_{h+1} \mid \widehat{c}_h, p_h, a_h).$$

Therefore, it suffices to learn $\mathbb{P}_h^{\pi_{1:h-1}^h,\mathcal{G}}(p_h|\widehat{c}_h)$ and $\mathbb{P}_h^{\pi_{1:h-1}^h,\mathcal{G}}(o_{h+1}|\widehat{c}_h,p_h,a_h)$. Similarly for $\widehat{r}^{\widetilde{\mathcal{M}}(\pi^{1:H})}$, it holds that

$$\widehat{r}_{i,h}^{\widetilde{\mathcal{M}}(\pi^{1:H})}(\widehat{c}_h,\gamma_h) = \sum_{p_h,a_h}\mathbb{P}_h^{\pi_{1:h-1}^h,\mathcal{G}}(p_h|\widehat{c}_h)\prod_{j=1}^n \gamma_{j,h}(a_{j,h}|p_{j,h})r_{i,h}^{\pi_{1:h-1}^h}(\widehat{c}_h,p_h,a_h),$$

where we define $r_{i,h}^{\pi_{1:h-1}^h}(\widehat{c}_h,p_h,a_h) := \mathbb{E}_{\pi_{1:h-1}^h}^{\mathcal{G}}[r_{i,h}(s_h,a_h)|\widehat{c}_h,p_h,a_h]$. 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.

Lemma 13. *Fix $\delta_1,\zeta_1,\zeta_2,\theta_1,\theta_2>0$. For Algorithm 5, suppose for all $h\in[H]$, $\pi^h\in\Delta(\Pi^{det})$ satisfies the conditions for π' of Lemma 12, then as long as N_0 in Algorithm 5 satisfies*

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

for some sufficiently large constant C, then with probability at least $1-\delta_1$, the following holds:

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

$$\sum_{p_h}\left|\mathbb{P}_h^{\widehat{\mathcal{M}}(\pi^{1:H})}(p_h|\widehat{c}_h)-\mathbb{P}_h^{\pi_{1:h-1}^h,\mathcal{G}}(p_h|\widehat{c}_h)\right|\leq\theta_1. \tag{E.13}$$

- *For all $h\in[H]$, $(\widehat{c}_h,p_h)\notin\mathcal{V}_{h,\zeta_2}^{low}(\pi^h)$, $a_h\in\mathcal{A}$, we have that*

$$\sum_{o_{h+1}}\left|\mathbb{P}_h^{\widehat{\mathcal{M}}(\pi^{1:H})}(o_{h+1}|\widehat{c}_h,p_h,a_h)-\mathbb{P}_h^{\pi_{1:h-1}^h,\mathcal{G}}(o_{h+1}|\widehat{c}_h,p_h,a_h)\right|\leq\theta_2, \tag{E.14}$$

$$\left|\widehat{r}_{i,h}^{\widehat{\mathcal{M}}(\pi^{1:H})}(\widehat{c}_h,p_h,a_h)-r_{i,h}^{\pi^{1:H}}(\widehat{c}_h,p_h,a_h)\right|\leq\theta_2. \tag{E.15}$$

We refer to the two bullets above as event $\mathcal{E}_1$.

Proof. We will prove Equation (E.13) first. Note that for any trajectory k of Algorithm 5, the distribution of p_h^k conditioned on $\widehat{c}_h^k$ is exactly $\mathbb{P}_h^{\pi_{1:h-1}^h,\mathcal{G}}(\cdot|\widehat{c}_h^k)$.

Now consider any $\widehat{c}_h\notin\mathcal{C}_{h,\zeta_1}^{\text{low}}(\pi^h)$. By the Chernoff bound, with probability at least $1-\exp(-\frac{\zeta_1 N_0}{8})$, there are at least $\frac{\zeta_1 N_0}{2}$ trajectories indexed by the set $\mathcal{K}^1\subseteq[N_0]$, such that for any $k\in\mathcal{K}^1$, $\text{Compress}_h(f_h(a_{1:h-1}^k,o_{1:h}^k))=\widehat{c}_h$. By the folklore theorem of learning a discrete probability distribution (Canonne, 2020), with probability at least $1-p'$, (E.13) holds as long as

$$\frac{\zeta_1 N_0}{2}\geq\frac{C(P_h+\log\frac{1}{p'})}{\theta_1^2}, \tag{E.16}$$

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.13) holds with probability at least

$$1-H\max_h\widehat{C}_h\exp(-\frac{\zeta_1 N_0}{8})-H\max_h\widehat{C}_h p'.$$

Now set $p' = \frac{\delta_1}{4H\max_h \widehat{C}_h}$ and it is easy to verify that (E.16) 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 $H\max_h \widehat{C}_h \exp(-\frac{\zeta_1 N_0}{8}) \leq \frac{\delta_1}{4}$. Therefore, we proved that with probability at least $1-\frac{\delta_1}{2}$, Equation (E.13) holds for all $h\in[H]$, and $\widehat{c}_h \notin \mathcal{C}_{h,\zeta_1}^{\mathrm{low}}(\pi^h)$.

Similarly, consider any trajectory k, the distribution of o_{h+1} conditioned on any $(\widehat{c}_h, p_h, a_h)$ is exactly $\mathbb{P}_h^{\pi^h_{1:h-1},\mathcal{G}}(\cdot\,|\,\widehat{c}_h, p_h, a_h)$. Now consider any $(\widehat{c}_h, p_h) \notin \mathcal{V}_{h,\zeta_2}^{\mathrm{low}}(\pi^h)$ and $a_h \in \mathcal{A}$. Note that due to the assumption on π^h that takes uniform random actions after step $h-L$, 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 the Chernoff bound, with probability at least $1-\exp(-\frac{\zeta_2 N_0}{8A})$, there are at least $\frac{\zeta_2 N_0}{2A}$ trajectories indexed by the set $\mathcal{K}^2 \subseteq [N_0]$, such that for any $k\in\mathcal{K}^2$, $\mathrm{Compress}_h(f_h(a_{1:h-1}^k, o_{1:h}^k)) = \widehat{c}_h, g_h(a_{1:h-1}^k, o_{1:h}^k) = p_h, a_h^k = a_h$. Again, with probability at least $1-p'$, (E.14) and (E.15) hold 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.14) and (E.15) hold 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.14) 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 conclude the proof. $\square$

With the previous lemma, the next step is to bound the two important quantities in Definition 7. In the following discussion, we will use the shorthand notation $\widetilde{\mathcal{M}}$ for $\widetilde{\mathcal{M}}(\pi^{1:H})$, and $\widehat{\mathcal{M}}$ for $\widehat{\mathcal{M}}(\pi^{1:H})$.

Lemma 14. *Under the event $\mathcal{E}_1$ in Lemma 13, for any $h\in[H]$, policy $\pi\in\Delta(\Pi^{\mathrm{det}})$, and prescription $\gamma_h\in\Gamma_h$, it holds that*

$$\mathbb{E}^{\mathcal{G}}_{a_{1:h-1},o_{1:h}\sim\pi}\sum_{z_{h+1}}\left|\mathbb{P}_h^{\widetilde{\mathcal{M}},z}(z_{h+1}\,|\,\widehat{c}_h,\gamma_h) - \mathbb{P}_h^{\widehat{\mathcal{M}},z}(z_{h+1}\,|\,\widehat{c}_h,\gamma_h)\right|$$
$$\leq \theta_1 + 2AP_h\frac{\zeta_2}{\zeta_1} + AP_h\theta_2 + \frac{A^{2\widehat{L}}O^{\widehat{L}}\zeta_1}{\phi} + \mathbf{1}[h>\widehat{L}]\cdot 2\cdot d_{\mathcal{S},h-\widehat{L}}^{\pi,\mathcal{G}}(\mathcal{U}^{\mathcal{G}}_{\phi,h-\widehat{L}}(\pi^h)), \tag{E.17}$$
$$\mathbb{E}^{\mathcal{G}}_{a_{1:h-1},o_{1:h}\sim\pi}\left|\widetilde{r}_{i,h}^{\widetilde{\mathcal{M}}}(\widehat{c}_h,\gamma_h) - \widetilde{r}_{i,h}^{\widehat{\mathcal{M}}}(\widehat{c}_h,\gamma_h)\right|$$
$$\leq \theta_1 + 2AP_h\frac{\zeta_2}{\zeta_1} + AP_h\theta_2 + \frac{A^{2\widehat{L}}O^{\widehat{L}}\zeta_1}{\phi} + \mathbf{1}[h>\widehat{L}]\cdot 2\cdot d_{\mathcal{S},h-\widehat{L}}^{\pi,\mathcal{G}}(\mathcal{U}^{\mathcal{G}}_{\phi,h-\widehat{L}}(\pi^h)). \tag{E.18}$$

Proof. It suffices to only consider $\pi\in\Pi^{\mathrm{det}}$, since if the statement holds for any $\pi\in\Pi^{\mathrm{det}}$, it will hold for any $\pi \in \Delta(\Pi^{\mathrm{det}})$ also. Under the event $\mathcal{E}_1$, consider any $\widehat{c}_h \notin \mathcal{C}^{\mathrm{low}}_{h,\zeta_1}(\pi^h)$ and $\gamma_h \in \Gamma_h$:

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

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

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

where the last step comes from Lemma 12. By noticing that after marginalization the total variation will not increase, we proved the first inequality.

Similarly, for the approximate reward, it holds that

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

Again by taking expectations over $\widehat{c}_h$, we proved the second inequality. □

Finally, we are ready to prove Theorem 4 by relating $\mathcal{G}$ and $\widehat{\mathcal{M}}(\pi^{1:H})$ through the intermediate $\widetilde{\mathcal{M}}(\pi^{1:H})$.

Proof of Theorem 4. 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 $\epsilon_r(\widehat{\mathcal{M}})$, it holds that

$$
\begin{aligned}
\epsilon_r(\widehat{\mathcal{M}})&=\max_{i,h}\max_{\pi\in\Pi^{\mathrm{det}},\gamma_h}\mathbb{E}_{a_{1:h-1},o_{1:h}\sim\pi}^{\mathcal{G}}|\mathbb{E}^{\mathcal{G}}[r_{i,h}(s_h,a_h)\mid c_h,\gamma_h]-\widehat{r}_{i,h}^{\widehat{\mathcal{M}}}(\widehat{c}_h,\gamma_h)| \\
&\le\max_{i,h}\max_{\pi\in\Pi^{\mathrm{det}},\gamma_h}\mathbb{E}_{a_{1:h-1},o_{1:h}\sim\pi}^{\mathcal{G}}|\mathbb{E}^{\mathcal{G}}[r_{i,h}(s_h,a_h)\mid c_h,\gamma_h]-\widetilde{r}_{i,h}^{\widetilde{\mathcal{M}}}(\widehat{c}_h,\gamma_h)| \\
&\quad+\max_{i,h}\max_{\pi\in\Pi^{\mathrm{det}},\gamma_h}\mathbb{E}_{a_{1:h-1},o_{1:h}\sim\pi}^{\mathcal{G}}|\widetilde{r}_{i,h}^{\widetilde{\mathcal{M}}}(\widehat{c}_h,\gamma_h)-\widehat{r}_{i,h}^{\widehat{\mathcal{M}}}(\widehat{c}_h,\gamma_h)| \\
&\le\epsilon_r(\pi^{1:H})+\epsilon_{apx}(\pi^{1:H},\widehat{L},\zeta_1,\zeta_2,\theta_1,\theta_2,\phi),
\end{aligned}
$$

where the last step comes from Lemma 14. Similarly, for $\epsilon_z(\widehat{\mathcal{M}})$, it holds that

$$
\begin{aligned}
\epsilon_z(\widehat{\mathcal{M}})&=\max_{h}\max_{\pi\in\Pi^{\mathrm{det}},\gamma_h}\mathbb{E}_{a_{1:h-1},o_{1:h}\sim\pi}^{\mathcal{G}}\|\mathbb{P}_h^{\mathcal{G}}(\cdot\,|\,c_h,\gamma_h)-\mathbb{P}_h^{\widehat{\mathcal{M}},z}(\cdot\,|\,\widehat{c}_h,\gamma_h)\|_1 \\
&\le\max_{h}\max_{\pi\in\Pi^{\mathrm{det}},\gamma_h}\mathbb{E}_{a_{1:h-1},o_{1:h}\sim\pi}^{\mathcal{G}}\|\mathbb{P}_h^{\mathcal{G}}(\cdot\,|\,c_h,\gamma_h)-\mathbb{P}_h^{\widetilde{\mathcal{M}},z}(\cdot\,|\,\widehat{c}_h,\gamma_h)\|_1 \\
&\quad+\max_{h}\max_{\pi\in\Pi^{\mathrm{det}},\gamma_h}\mathbb{E}_{a_{1:h-1},o_{1:h}\sim\pi}^{\mathcal{G}}\|\mathbb{P}_h^{\widetilde{\mathcal{M}},z}(\cdot\,|\,c_h,\gamma_h)-\mathbb{P}_h^{\widehat{\mathcal{M}},z}(\cdot\,|\,\widehat{c}_h,\gamma_h)\|_1 \\
&\le\epsilon_z(\pi^{1:H})+\epsilon_{apx}(\pi^{1:H},\widehat{L},\zeta_1,\zeta_2,\theta_1,\theta_2,\phi),
\end{aligned}
$$

where the last step again comes from Lemma 14. Therefore, with Lemma 2 and Theorem 2, we proved Theorem 8. □

E.6 Proof of Theorem 5

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 common information $\widehat{c}_h$ using finite memory. Similarly, here we will also show how to construct $\widehat{c}_h$ with finite memory so that $\widetilde{\mathcal{M}}(\pi^{1:H})$ is a good approximation of $\mathcal{G}$. In the following discussions, we shall use another important policy-dependent approximate belief $\widetilde{\boldsymbol{b}}_h^{\pi}(\cdot) := \boldsymbol{b}_h^{\mathrm{apx},\mathcal{G}}(\cdot; d_{\mathcal{S},h-L}^{\pi,\mathcal{G}})$. We first introduce the following important lemmas.

Lemma 15. *There is a constant $C \geq 1$ so that the following holds. If Assumption 2 holds, then for any $\epsilon, \phi > 0, L \in \mathbb{N}$ so that $L \geq C\gamma^{-4}\log(\frac{1}{\epsilon\phi})$, it holds that for any policies $\pi, \pi' \in \Delta(\Pi^{\mathrm{det}})$,*

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

Furthermore, for any finite domain Y, conditional probability $q(y \mid s)$, and the posterior update operator $F^q : \Delta(\mathcal{S}) \to \Delta(\mathcal{S})$ as defined in Lemma 9, it holds that

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

Proof. It directly follows from our Theorem 10, and Lemma 12.2 of Golowich et al. (2022a). □

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

Lemma 16. *Given $\alpha, \beta > 0$, $\widehat{L} \geq C\frac{\log(HSO/(\alpha\gamma))}{\gamma^4}$, and $\phi = \frac{\alpha\gamma^2}{C^3H^{10}S^5O^4}$ for some constant $C > 0$. There exists an algorithm BaSeCAMP (Algorithm 3 of Golowich et al. (2022a)) 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 $\pi^{h,j} \in \Delta(\Pi^{det})$ and $\pi_{h'}^{h,j} = \mathrm{Unif}(\mathcal{A})$ for $h' \geq h - \widehat{L}, j \in [K]$. 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 $\pi \in \Pi^{\mathrm{det}}$:*

$$d_{\mathcal{S},h-\widehat{L}}^{\pi,\mathcal{G}}(\mathcal{U}_{\phi,h-\widehat{L}}^{\mathcal{G}}(\pi^{h,j^\star})) \leq \frac{\alpha}{CH^2}.$$

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

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

Corollary 4. *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 (Algorithm 3 of Golowich et al. (2022a)) 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 $\pi^{h,j} \in \Delta(\Pi^{\mathrm{det}})$ and $\pi_{h'}^{h,j} = \mathrm{Unif}(\mathcal{A})$ for $h \in [H]$, $h' \geq h - L$, $j \in [K]$. 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 $\pi' \in \Delta(\Pi^{\rm det})$:

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

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

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

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

Proof. Let $\alpha = \frac{CH^2\epsilon}{2}$, $\delta_2 = \beta$, and $L \geq \max\{C\frac{\log(\frac{1}{\epsilon\phi})}{\gamma^4}, C\frac{\log(HSO/(\alpha\gamma))}{\gamma^4}\}$. Combining Lemmas 15 and 16 leads to the conclusion. □

In the discussion thereafter, we will use $\widetilde{\mathcal{M}}$ for $\widetilde{\mathcal{M}}(\pi^{1:H,j^\star})$ and $\widehat{\mathcal{M}}$ for $\widehat{\mathcal{M}}(\pi^{1:H,j^\star})$, and $\widehat{r}_{i,h}$ for $\widehat{r}_{i,h}^{j^\star}$ 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 $\{\pi^{\star,j}\}_{j=1}^K$, which are generated by running Algorithm 3 on the candidate models $\{\widehat{\mathcal{M}}(\pi^{1:H,j})\}_{j\in[K]}$. The policy evaluation and selection algorithm is described in Algorithm 7.

Lemma 17. *For Algorithm 7, 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*

$$\left|V_{i,1}^{\pi,\mathcal{G}}(\emptyset) - V_{i,1}^{\pi,\widehat{\mathcal{M}}(\pi^{1:H,j^\star})}(\emptyset)\right| \leq \epsilon.$$

If $N_2 \geq C\frac{H^2\log\frac{K^2n}{\delta_3}}{\epsilon^2}$ for some constant $C > 0$, then with probability at least $1-\delta_3$, the following event $\mathcal{E}_3$ holds

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

Proof. For NE/CCE, note that $V_{i,1}^{\pi_i^{\star,j,m}\times\pi_{-i}^{\star,j},\widehat{\mathcal{M}}(\pi^{1:H,m})}(\emptyset) \geq \max_{\pi_i} V_{i,1}^{\pi_i\times\pi_{-i}^{\star,j},\widehat{\mathcal{M}}(\pi^{1:H,m})}(\emptyset) - H\epsilon_e$ according to Corollary 1 for $m \in [K]$. By the concentration bound on the relationship between the accumulated rewards and the value function for all policies $\pi^{\star,j}, \pi_i^{\star,j,m} \times \pi_{-i}^{\star,j}$, and further a union bound over all $i \in [n]$, $j \in [K]$, and $m \in [K]$, with probability at least $1-\delta_3$, the following event $\mathcal{E}_3$ holds for any $i \in [n], j \in [K], m \in [K]$:

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

In the following proof, we will assume the previous event holds. Define $m_{i,j}^\star \in \arg\max_m R_i^{j,m}$. 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 Corollary 1 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 $\pi^{\star,j^\star}$, 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 $\widehat{j} \in \arg\min_j\left(\max_i \max_m (R_i^{j,m} - R_i^j)\right)$, 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, note that

$$V_{i,1}^{\pi_i^{\star,j,m}\odot\pi_{-i}^{\star,j},\widehat{\mathcal{M}}(\pi^{1:H,m})}(\emptyset) \ge \max_{\phi_i} V_{i,1}^{(\phi_i\diamond\pi_i^{\star,j})\odot\pi_{-i}^{\star,j},\widehat{\mathcal{M}}(\pi^{1:H,m})}(\emptyset) - H\epsilon_e.$$

Similarly, by a concentration bound and then a union bound, with probability at least $1-\delta_3$, the following event $\mathcal{E}_3$ holds for any $i \in [n], j \in [K], m \in [K]$:

$$\left|R_i^j - V_{i,1}^{\pi^{\star,j},\mathcal{G}}(\emptyset)\right| \le \epsilon, \qquad \left|R_i^{j,m} - V_{i,1}^{\pi_i^{\star,j,m}\odot\pi_{-i}^{\star,j},\mathcal{G}}(\emptyset)\right| \le \epsilon.$$

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 strategy modification with respect to $\pi_{-i}^{\star,j}$. Note that for any $i \in [n], j \in [K]$:

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

On the other hand,

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

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

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

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 9. Before diving into the examples in Appendix B, the proof for the first part of Theorem 8 follows from the fact that both the computation and sample complexities depend on $\max_h C_h$ and $\max_h P_h$. Therefore, if we can find $\pi^{1:H}$ and Compress_h for $h \in [H]$ such that the relevant errors are minimized while $\max_h C_h$ and $\max_h P_h$ are of quasi-polynomial size, then there exists a quasi-polynomial sample and time algorithm learning ϵ-NE if $\mathcal{G}$ is zero-sum or cooperative and ϵ-CE/CCE if $\mathcal{G}$ is general-sum. In the following discussion, we will see the sample complexity of our algorithm instantiated with specific information structures.

One-step delayed information sharing. In this case, the information structure gives $c_h = \{a_{1:h-1}, o_{1:h-1}\}$, $p_{i,h} = \{o_{i,h}\}$, $z_{h+1} = \{o_h, a_h\}$. 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 $\pi^{1:H}$, where $\pi^h \in \Delta(\Pi^{\text{det}})$ for $h \in [H]$, it is direct to verify that

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

where we recall the definition of $\widetilde{\boldsymbol{b}}_h^{\pi^h}$ in Appendix E.6. Meanwhile, according to Definition 10, it is direct to verify that $\widehat{L} = L$. Hereafter in the proof, we use $\widetilde{\mathcal{M}}$ to denote $\widetilde{\mathcal{M}}(\pi^{1:H,j^\star})$ for short. Therefore, we conclude that if $L \geq C\frac{\log(HSO/(\epsilon\gamma))}{\gamma^4}$, by a union bound of the high probability event $\mathcal{E}_1$ in Lemma 13, $\mathcal{E}_2$ in Corollary 4, and $\mathcal{E}_3$ in Lemma 17, with probability at least $1-\delta_1-\delta_2-\delta_3$, it holds that for any $i \in [n]$

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

and moreover

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

According to the choice $\pi^{1:H,j^\star}$, it holds that by Corollary 4

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

Therefore, for any $\alpha, \delta > 0$, setting $\epsilon = \frac{\alpha}{200(H+1)^2}$, $\theta_1 = \frac{\alpha}{200(H+1)^2 O}$, $\zeta_2 = \zeta_1^2$, $\theta_2 = \frac{\alpha}{200(H+1)^2 A \max_h P_h}$, $\zeta_1 = \min\left\{\frac{\alpha\phi}{200(H+1)^2 A^{2L} O^L}, \frac{\alpha}{400(H+1)^2 A \max_h P_h}\right\}$, $\phi = \frac{\epsilon\gamma^2}{C^2 H^8 S^5 O^4}$, $\epsilon_e = \frac{\alpha}{200H}$, $\delta_1 = \delta_2 = \delta_3 = \frac{\delta}{3}$, $\widetilde{\mathcal{M}}(\pi^{1:H,j^\star})$ 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 $|V_{i,1}^{\pi,\mathcal{G}}(\emptyset) - V_{i,1}^{\pi,\widehat{\mathcal{M}}(\pi^{1:H,j^\star})}(\emptyset)| \leq \frac{15\alpha}{200}$ for any policy $\pi \in \Pi$ by Lemma 2. By Lemma 17, 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 complexities of our algorithm.

Theorem 16. *Let $\alpha, \delta, \gamma > 0$. Algorithm 9 given a γ-observable POSG of one-step delayed information sharing structure outputs an α-NE if the POSG is zero-sum or cooperative, or α-CE/CCE if the POSG is general-sum, with probability at least $1-\delta$, with time and sample complexities bounded by $(AO)^{C\gamma^{-4}\log\frac{SHO}{\gamma\alpha}} \log\frac{1}{\delta}$ for some universal constant $C > 0$.*

Proof. Recall that $\widehat{C}_h \leq (OA)^L$, $P_h \leq O$, $N_0 = \max\left\{\frac{C(\max_h P_h + \log\frac{4H\max_h \widehat{C}_h}{\delta_1})}{\zeta_1\theta_1^2}, \frac{CA(O+\log\frac{4H\max_h(\widehat{C}_h P_h A)}{\delta_1})}{\zeta_2\theta_2^2}\right\}$, $N_1 = (OA)^L \log(\frac{1}{\delta_2})$, and $N_2 = C\frac{H^2\log\frac{K^2 n}{\delta_3}}{\epsilon^2}$ for some constant $C > 0$, and we have set $\delta_1 = \delta_2 = \delta_3 = \frac{\delta}{3}$. 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, since our algorithm only calls the BaSeCAMP and our planning algorithm polynomial number of times, the 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 $c_h = \{o_{1,1:h}, o_{2,1:h-d}, a_{1,1:h-1}\}$, $p_{1,h} = \emptyset$, $p_{2,h} = \{o_{2,h-d+1:h}\}$. Fix some $L > 0$, 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 $\pi^{1:H}$, where $\pi^h \in \Delta(\Pi^{\mathrm{det}})$, following exactly the same derivation as in Appendix E.4, it holds that

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

Meanwhile, it is direct to verify that $\widehat{L} = L + d$ by Definition 10. Therefore, we conclude that if $L \geq C\frac{\log(HSO/(\epsilon\gamma))}{\gamma^4}$, by a union bound of the high probability event $\mathcal{E}_1$ in Lemma 13, $\mathcal{E}_2$ in Corollary 4, and $\mathcal{E}_3$ in Lemma 17, with probability at least $1 - \delta_1 - \delta_2 - \delta_3$, it holds that for any $i \in [n]$:

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

and moreover

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

According to the choice $\pi^{1:H,j^\star}$, it holds that by Corollary 4

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

Therefore, for any $\alpha, \delta > 0$, setting $\epsilon = \frac{\alpha}{200(H+1)^2}$, $\theta_1 = \frac{\alpha}{200(H+1)^2 O}$, $\zeta_2 = \zeta_1^2$, $\theta_2 = \frac{\alpha}{200(H+1)^2 A \max_h P_h}$, $\zeta_1 = \min\left\{\frac{\alpha\phi}{200(H+1)^2 A^{2(L+d)} O^{L+d}}, \frac{\alpha}{400(H+1)^2 A \max_h P_h}\right\}$, $\phi = \frac{\epsilon\gamma^2}{C^2 H^8 S^5 O^4}$, $\epsilon_e = \frac{\alpha}{200H}$, $\delta_1 = \delta_2 = \delta_3 = \frac{\delta}{3}$, $\widetilde{\mathcal{M}}(\pi^{1:H,j^\star})$ 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 $|V_{i,1}^{\pi,\mathcal{G}}(\emptyset) - V_{i,1}^{\pi,\widehat{\mathcal{M}}(\pi^{1:H,j^\star})}(\emptyset)| \leq \frac{15\alpha}{200}$ for any policy $\pi \in \Pi$ by Lemma 2. By Lemma 17, 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 complexities of our algorithm.

Theorem 17. *Let $\alpha, \delta, \gamma > 0$. Algorithm 9 given a γ-observable POSG of state controlled by one controller with asymmetric delay sharing outputs an α-NE if the POSG is zero-sum or cooperative, or α-CE/CCE if the POSG is general-sum, with probability at least $1 - \delta$, with time and sample complexities bounded by $(OA)^{C(\gamma^{-4}\log\frac{SHO}{\gamma\alpha} + d)} \log\frac{1}{\delta}$ for some universal constant $C > 0$.*

Proof. Recall that $\widehat{C}_h \le (AO)^L$, $P_h \le (AO)^d$, $N_0 = \max\left\{\frac{C(\max_h P_h + \log\frac{4H\max_h \widehat{C}_h}{\delta_1})}{\zeta_1\theta_1^2}, \frac{CA(O+\log\frac{4H\max_h(\widehat{C}_h P_h)A}{\delta_1})}{\zeta_2\theta_2^2}\right\}$, $N_1 = (OA)^{\widehat{L}}\log(\frac{1}{\delta_2})$, and $N_2 = C\frac{H^2\log\frac{K^2 n}{\delta_3}}{\epsilon^2}$ for some constant $C>0$, and we have set $\delta_1=\delta_2=\delta_3=\frac{\delta}{3}$. 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, the time complexity is also bounded by $(OA)^{C(\gamma^{-4}\log\frac{SHO}{\gamma\alpha}+d)}\log\frac{1}{\delta}$. $\square$

Information sharing with one-directional-one-step delay. For this case, we have $c_h = \{o_{1,1:h}, o_{2,1:h-1}, a_{1:h-1}\}$, $p_{1,h} = \emptyset$, $p_{2,h} = \{o_{2,h}\}$, 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 $\pi^{1:H}$, where $\pi^h \in \Delta(\Pi^{\mathrm{det}})$ for $h\in[H]$, it is easy to verify that

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

where $\mathbb{P}_h(o_{2,h}\mid s_h, o_{1,h}) = \frac{\mathbb{O}_h(o_{1,h}, o_{2,h}\mid s_h)}{\sum_{o'_{2,h}}\mathbb{O}_h(o_{1,h}, o'_{2,h}\mid s_h)}$. Furthermore, it is direct to verify that $\widehat{L}=L$. Therefore, we conclude that if $L \ge C\frac{\log(HSO/(\epsilon\gamma))}{\gamma^4}$, by a union bound of the high probability event $\mathcal{E}_1$ in Lemma 13, $\mathcal{E}_2$ in Corollary 4, and $\mathcal{E}_3$ in Lemma 17, with probability at least $1-\delta_1-\delta_2-\delta_3$, it holds that for any $i\in[n]$:

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

Moreover, we have

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

According to the choice $\pi^{1:H,j^\star}$, it holds that by Corollary 4:

$$\max_h\max_{\pi\in\Pi^{\mathrm{det}}}\mathbf{1}[h>L]\cdot 6\cdot d_{\mathcal{S},h-L}^{\pi,\mathcal{G}}\left(\mathcal{U}_{\phi,h-L}^{\mathcal{G}}\left(\pi^{h,j^\star}\right)\right) \le 6\epsilon.$$

Therefore, for any $\alpha,\delta>0$, setting $\epsilon = \frac{\alpha}{200(H+1)^2}$, $\theta_1 = \frac{\alpha}{200(H+1)^2 O}$, $\zeta_2 = \zeta_1^2$, $\theta_2 = \frac{\alpha}{200(H+1)^2 A\max_h P_h}$, $\zeta_1 = \min\left\{\frac{\alpha\phi}{200(H+1)^2A^{2L}O^L}, \frac{\alpha}{400(H+1)^2A\max_h P_h}\right\}$, $\phi = \frac{\epsilon\gamma^2}{C^2H^8S^5O^4}$, $\epsilon_e = \frac{\alpha}{200H}$, $\delta_1=\delta_2=\delta_3=\frac{\delta}{3}$, $\widetilde{\mathcal{M}}(\pi^{1:H,j^\star})$ is an (ϵ_r,ϵ_z)-expected-approximate common information model of $\mathcal{G}$, where $\epsilon_r,\epsilon_z \le \frac{14\alpha}{200(H+1)^2}$. This leads to that $\pi^{\star,j^\star}$ is a $\frac{15\alpha}{200}$-NE/CE/CCE, and $|V_{i,1}^{\pi,\mathcal{G}}(\emptyset) - V_{i,1}^{\pi,\widetilde{\mathcal{M}}(\pi^{1:H,j^\star})}(\emptyset)| \le \frac{15\alpha}{200}$ for any policy $\pi\in\Pi$ by Lemma

2. By Lemma 17, 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 complexities of our algorithm.

Theorem 18. *Let $\alpha, \delta, \gamma > 0$. Algorithm 9 given a γ-observable POSG of one-directional-one-step delayed information sharing structure outputs an α-NE if the POSG is zero-sum or cooperative, or α-CE/CCE if the POSG is general-sum, with probability at least $1-\delta$, with time and sample complexities bounded by $(AO)^{C\gamma^{-4}\log\frac{SHO}{\gamma\alpha}}\log\frac{1}{\delta}$ for some universal constant $C>0$.*

Proof. Recall that $\widehat{C}_h \leq (OA)^L$, $P_h \leq O$, $N_0 = \max\left\{\frac{C(\max_h P_h + \log\frac{4H\max_h \widehat{C}_h}{\delta_1})}{\zeta_1\theta_1^2}, \frac{CA(O+\log\frac{4H\max_h(\widehat{C}_h P_h A)}{\delta_1})}{\zeta_2\theta_2^2}\right\}$, $N_1 = (OA)^L\log(\frac{1}{\delta_2})$, and $N_2 = C\frac{H^2\log\frac{K^2 n}{\delta_3}}{\epsilon^2}$ for some constant $C>0$, and we have set $\delta_1 = \delta_2 = \delta_3 = \frac{\delta}{3}$. 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, the 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 gives that $c_h = \{o_{1:h-d}\}$, $p_{i,h} = \{o_{i,h-d+1:h}\}$, and $z_{h+1} = \{o_{h-d+1}\}$. Fix a $L>0$, the approximate common information is $\widehat{c}_h = \{o_{h-d-L+1:h-d}\}$. For any policy $\pi^{1:H}$, where $\pi^h \in \Delta(\Pi^{\mathrm{det}})$ for $h\in[H]$, it is easy to verify that

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

Furthermore, it is direct to verify that $\widehat{L} = L+d$ by Definition 10. Therefore, we conclude that if $L \geq C\frac{\log(HSO/(\epsilon\gamma))}{\gamma^4}$, by a union bound of the high probability event $\mathcal{E}_1$ in Lemma 13, $\mathcal{E}_2$ in Corollary 4, and $\mathcal{E}_3$ in Lemma 17, with probability at least $1-\delta_1-\delta_2-\delta_3$, it holds that for any $i\in[n]$:

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

Moreover, we also have

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

According to the choice $\pi^{1:H,j^\star}$, it holds that by Corollary 4:

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

Therefore, for any $\alpha,\delta>0$, setting $\epsilon = \frac{\alpha}{200(H+1)^2}$, $\theta_1 = \frac{\alpha}{200(H+1)^2 O}$, $\zeta_2 = \zeta_1^2$, $\theta_2 = \frac{\alpha}{200(H+1)^2 A\max_h P_h}$, $\zeta_1 = \min\left\{\frac{\alpha\phi}{200(H+1)^2 A^{2(L+d)}O^{L+d}}, \frac{\alpha}{400(H+1)^2 A\max_h P_h}\right\}$, $\phi = \frac{\epsilon\gamma^2}{C^2H^8S^5O^4}$, $\epsilon_e = \frac{\alpha}{200H}$, $\delta_1 = \delta_2 = \delta_3 = \frac{\delta}{3}$, $\widetilde{\mathcal{M}}(\pi^{1:H,j^\star})$

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 $|V_{i,1}^{\pi,\mathcal{G}}(\emptyset) - V_{i,1}^{\pi,\widehat{\mathcal{M}}(\pi^{1:H,j^\star})}(\emptyset)| \leq \frac{15\alpha}{200}$ for any policy π by Lemma 2. By Lemma 17,

$$\text{NE/CE/CCE-gap}(\pi^{\star,\widehat{j}}) \leq \text{NE/CE/CCE-gap}(\pi^{\star,j^\star}) + \frac{91\alpha}{200} \leq \alpha.$$

Finally, we are ready to analyze the computation and sample complexity of our algorithm.

Theorem 19. *Let $\alpha, \delta, \gamma > 0$. Algorithm 9 given a γ-observable POSG of uncontrolled state process and delayed information sharing structure outputs an α-NE if the POSG is zero-sum or cooperative, or α-CE/CCE if the POSG is general-sum, with probability at least $1-\delta$, with time and sample complexities bounded by $(OA)^{C(\gamma^{-4}\log\frac{SHO}{\gamma\alpha}+d)}\log\frac{1}{\delta}$ for some universal constant $C > 0$.*

Proof. Recall that $\widehat{C}_h \leq O^L$, $P_h \leq O^d$, $N_0 = \max\left\{\frac{C(\max_h P_h + \log\frac{4H\max_h \widehat{C}_h}{\delta_1})}{\zeta_1\theta_1^2}, \frac{CA(O+\log\frac{4H\max_h(\widehat{C}_h P_h)A}{\delta_1})}{\zeta_2\theta_2^2}\right\}$, $N_1 = (OA)^{\widehat{L}}\log(\frac{1}{\delta_2})$, and $N_2 = C\frac{H^2\log\frac{K^2 n}{\delta_3}}{\epsilon^2}$ for some constant $C > 0$, and we have set $\delta_1 = \delta_2 = \delta_3 = \frac{\delta}{3}$. 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, the 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 $L > 0$, we construct the approximate common information as $\widehat{c}_h = \{o_{h-L+1:h}, a_{h-L:h-1}\}$. For any $\pi^{1:H}$, where $\pi^h \in \Delta(\Pi^{\text{det}})$ for $h \in [H]$, it is easy to verify that

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

Meanwhile, it is direct to verify that $\widehat{L} = L$ by Definition 10. Therefore, we conclude that if $L \geq C\frac{\log(HSO/(\epsilon\gamma))}{\gamma^4}$, by a union bound of the high probability event $\mathcal{E}_1$ in Lemma 13, $\mathcal{E}_2$ in Corollary 4, and $\mathcal{E}_3$ in Lemma 17, with probability at least $1 - \delta_1 - \delta_2 - \delta_3$, it holds that for any $i \in [n]$:

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

Moreover, we have

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

According to the choice $\pi^{1:H,j^\star}$, it holds that by Corollary 4

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

Therefore, for any $\alpha, \delta > 0$, setting $\epsilon = \frac{\alpha}{200(H+1)^2}$, $\theta_1 = \frac{\alpha}{200(H+1)^2 O}$, $\zeta_2 = \zeta_1^2$, $\theta_2 = \frac{\alpha}{200(H+1)^2 A \max_h P_h}$, $\zeta_1 = \min\left\{\frac{\alpha\phi}{200(H+1)^2 A^{2L} O^L}, \frac{\alpha}{400(H+1)^2 A \max_h P_h}\right\}$, $\phi = \frac{\epsilon\gamma^2}{C^2 H^8 S^5 O^4}$, $\epsilon_e = \frac{\alpha}{200H}$, $\delta_1 = \delta_2 = \delta_3 = \frac{\delta}{3}$, $\widetilde{\mathcal{M}}(\pi^{1:H,j^\star})$ 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 $|V_{i,1}^{\pi,\mathcal{G}}(\emptyset) - V_{i,1}^{\pi,\widehat{\mathcal{M}}(\pi^{1:H,j^\star})}(\emptyset)| \leq \frac{15\alpha}{200}$ for any policy $\pi \in \Pi$ by Lemma 2. By Lemma 17, 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 complexities of our algorithm.

Theorem 20. *Let $\alpha, \delta, \gamma > 0$. Algorithm 9 given a γ-observable POSG of symmetric information sharing structure outputs an α-NE if the POSG is zero-sum or cooperative, or α-CE/CCE if the POSG is general-sum, with probability at least $1-\delta$, with time and sample complexities bounded by $(AO)^{C\gamma^{-4}\log\frac{SHO}{\gamma\alpha}} \log\frac{1}{\delta}$ for some universal constant $C > 0$.*

Proof. Recall that $\widehat{C}_h \leq (OA)^L$, $P_h = 1$, $N_0 = \max\left\{\frac{C(\max_h P_h + \log\frac{4H\max_h \widehat{C}_h}{\delta_1})}{\zeta_1\theta_1^2}, \frac{CA(O + \log\frac{4H\max_h(\widehat{C}_h P_h)A}{\delta_1})}{\zeta_2\theta_2^2}\right\}$, $N_1 = (OA)^L \log(\frac{1}{\delta_2})$, and $N_2 = C\frac{H^2\log\frac{K^2 n}{\delta_3}}{\epsilon^2}$ for some constant $C > 0$, and we have set $\delta_1 = \delta_2 = \delta_3 = \frac{\delta}{3}$. 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, the time complexity is also bounded by $(OA)^{C(\gamma^{-4}\log\frac{SHO}{\gamma\alpha})}\log\frac{1}{\delta}$. $\square$

E.7 Missing details in Section 5

Now we prove Proposition 2, where the hardness follows from the hardness of the one-step Dec-POMDP in Proposition 4.

Proof of Proposition 2. Note that for Equation (5.1), if we take the underlying Dec-POMDP $\mathcal{G}$ to be $H = 1$, $n = 2$ without any information-sharing, and the approximate belief is constructed to be the ground-truth belief of the underlying Dec-POMDP $\mathcal{G}$, the optimal prescription solved by Equation (5.1) is then exactly the optimal policy of the underlying $\mathcal{G}$. By the hardness from Proposition 4, we conclude that solving Equation (5.1) is also NP-hard. $\square$

Proposition 11. *Given any approximate common information model $\mathcal{M}$ that is consistent with a belief $\{\mathbb{P}_{h'}^{\mathcal{M},c}(s_{h'}, p_{h'} \mid \widehat{c}_{h'})\}_{h' \in [H]}$, if **Condition 1** holds, we have for any $h \in [H]$, $\widehat{c}_h \in \widehat{\mathcal{C}}_h$, $\gamma_h \in \Gamma_h$*

$$Q_h^{\star,\mathcal{M}}(\widehat{c}_h, \gamma_h) = \sum_{j \in [n]} U_{j,h}(\widehat{c}_h, \gamma_{j,h}),$$

for some functions $\{U_{j,h}\}_{j \in [n]}$. Correspondingly, Equation (5.1) can be solved exactly in time complexity poly(S, A, P_h).

Proof of Proposition 11. By the definition of $Q_h^{\star,\mathcal{M}}(\widehat{c}_h, \gamma_h)$ and Definition 8, it holds that

$$
\begin{aligned}
&Q_h^{\star,\mathcal{M}}(\widehat{c}_h, \gamma_h)\\
&= \sum_{s_h, p_h, a_h, s_{h+1}, o_{h+1}} \mathbb{P}_h^{\mathcal{M},c}(s_h, p_h \mid \widehat{c}_h) \prod_{i=1}^{n} \gamma_{i,h}(a_{i,h} \mid p_{i,h}) \mathbb{T}_h(s_{h+1} \mid s_h, a_{\mathrm{ctt}(h),h})\\
&\quad \mathbb{O}_{h+1}(o_{h+1} \mid s_{h+1}) \Big[r_h(s_h, a_h) + V_{h+1}^{\star,\mathcal{M}}(\widehat{c}_{h+1}) \Big]\\
&= \sum_{j \neq \mathrm{ctt}(h)} \sum_{s_h, p_h, a_{j,h}} \mathbb{P}_h^{\mathcal{M},c}(s_h, p_h \mid \widehat{c}_h) \gamma_{j,h}(a_{j,h} \mid p_{j,h}) \Big[r_{j,h}(s_h, a_{j,h}) \Big] +\\
&\quad \sum_{s_h, p_h, a_{\mathrm{ctt}(h),h}, s_{h+1}, o_{h+1}} \mathbb{P}_h^{\mathcal{M},c}(s_h, p_h \mid \widehat{c}_h) \gamma_{\mathrm{ctt}(h),h}(a_{\mathrm{ctt}(h),h} \mid p_{\mathrm{ctt}(h),h}) \mathbb{T}_h(s_{h+1} \mid s_h, a_{\mathrm{ctt}(h),h}) \mathbb{O}_{h+1}(o_{h+1} \mid s_{h+1})\\
&\qquad\qquad \times \Big[r_{\mathrm{ctt}(h),h}(s_h, a_{\mathrm{ctt}(h),h}) + V_{h+1}^{\star,\mathcal{M}}(\widehat{c}_{h+1}) \Big]\\
&:= \sum_{j \in [n]} U_{j,h}(\widehat{c}_h, \gamma_{j,h})
\end{aligned}
$$

where the last step is due to the assumption that that $\widehat{c}_{h+1} = \widehat{\phi}_{h+1}(\widehat{c}_h, z_{h+1})$ and $z_{h+1} = \chi_{h+1}(p_h, a_{\mathrm{ctt}(h),h}, o_{h+1})$. Now, to solve Equation (5.1), we only need to optimize w.r.t. each $\gamma_{j,h}$ for $j \in [n]$ *individually*, which is a linear program with the constraint set of $\gamma_{j,h}$ to be a concatenation of simplex by Proposition 8. Hence, Equation (5.1) can be solved even *exactly* in time complexity $\texttt{poly}(S, A, P_h)$. $\square$

Proposition 12. *Suppose **Condition 2** holds, Algorithm 10 returns $\gamma_{1:n,h}^{\star}$ such that*

$$
\gamma_{1:n,h}^{\star} \in \arg\max_{\gamma_{1,h}, \cdots, \gamma_{n,h}} Q_h^{\star,\mathcal{M}}(\widehat{c}_h, \gamma_{1,h}, \cdots, \gamma_{n,h}),
$$

with time complexity $\texttt{poly}(P_h, A, S)$.

Proof of Proposition 12. We slightly abuse our notation for the $Q_h^{\star,\mathcal{M}}$ as below to define for any $u_i \in \mathcal{U}_i := \{(\times_{j=1}^{i} \mathcal{P}_{j,h}) \times (\times_{j=1}^{i-1} \mathcal{A}_j) \to \Delta(\mathcal{A}_i)\}$ and $i \in [n]$ that

$$
\begin{aligned}
&Q_h^{\star,\mathcal{M}}(\widehat{c}_h, u_1, \cdots, u_n)\\
&:= \sum_{s_h, p_h, a_h, s_{h+1}, o_{h+1}} \mathbb{P}_h^{\mathcal{M},c}(s_h, p_h \mid \widehat{c}_h) \prod_{i=1}^{n} u_i(a_{i,h} \mid p_{1:i,h}, a_{1:i-1,h}) \mathbb{T}_h(s_{h+1} \mid s_h, a_h) \mathbb{O}_{h+1}(o_{h+1} \mid s_{h+1}) \Big[r_h(s_h, a_h) + V_{h+1}^{\star,\mathcal{M}}(\widehat{c}_{h+1}) \Big].
\end{aligned}
$$

By the standard result of value iteration for POMDPs, we have that $u_{1:n}^{\star}$ is an optimal policy for the POMDP $\widehat{\mathcal{P}}(n)$ in the sense that

$$
Q_h^{\star,\mathcal{M}}(\widehat{c}_h, u_1^{\star}, \cdots, u_n^{\star}) = \max_{\{u_i \in \mathcal{U}_i\}_{i \in [n]}} Q_h^{\star,\mathcal{M}}(\widehat{c}_h, u_1, \cdots, u_n) \geq \max_{\{\gamma_{i,h} \in \Gamma_{i,h}\}_{i \in [n]}} Q_h^{\star,\mathcal{M}}(\widehat{c}_h, \gamma_{i,h}, \cdots, \gamma_{n,h}),
$$

where the inequality comes from the fact that any $\gamma_{i,h} \in \Gamma_{i,h}$ can be realized by an equivalent $u_i \in \mathcal{U}_i$ such that the value is the same. Meanwhile, due to the nested information-sharing structure, for any $p_h \in \mathcal{P}_h$, it holds that $u_{1:n}^{\star}$ and $\gamma_{1:n,h}^{\star}$ outputs the same action deterministically according to the second for-loop of Algorithm 10. Hence, we conclude that

$$
Q_h^{\star,\mathcal{M}}(\widehat{c}_h, \gamma_{1,h}^{\star}, \cdots, \gamma_{n,h}^{\star}) = Q_h^{\star,\mathcal{M}}(\widehat{c}_h, u_1^{\star}, \cdots, u_n^{\star}),
$$

which further concludes that $\gamma^\star_{1:n,h}$ returned by Algorithm 10 is an exact solution of Equation (5.1). Finally, the time complexity scales with the size of the history space of $\widehat{\mathcal{P}}(n)$, which is $\prod_{i=1}^n P_{i,h}A_{i,h} = AP_h$. The additional polynomial dependency on S comes from computing the posterior distribution for the initialization step in Algorithm 10. $\square$

Proposition 13. *Suppose **Condition 3** holds. For each $h\in[H]$, there exist n functions $\{F_{i,h}\}_{i\in[n]}$ such that*

$$Q_h^{\star,\mathcal{M}}(\widehat{c}_h,\gamma_{1,h},\cdots,\gamma_{n,h}) = \sum_{i=1}^n F_{i,h}(\widehat{c}_{i,h},\gamma_{i,h}).$$

Correspondingly, Equation (5.1) can be solved in time $\sum_{i\in[n]}\mathrm{poly}(S_i,A_i,P_{i,h})$.

Proof of Proposition 13. We prove our result by backward induction on h. Obviously, it holds for $h=H+1$. Now suppose the proposition holds for $h+1$. For step h, it holds that

$$\begin{aligned}
&Q_h^{\star,\mathcal{M}}(\widehat{c}_h,\gamma_{1,h},\cdots,\gamma_{n,h})\\
&=\sum_{s_h,p_h,a_h,s_{h+1},o_{h+1}}\mathbb{P}_h^{\mathcal{M},c}(s_h,p_h\,|\,\widehat{c}_h)\prod_{j=1}^n\gamma_{j,h}(a_{j,h}\,|\,p_{j,h})\mathbb{T}_h(s_{h+1}\,|\,s_h,a_h)\\
&\qquad\qquad\mathbb{O}_{h+1}(o_{h+1}\,|\,s_{h+1})\left[\sum_{i=1}^n r_{i,h}(s_{i,h},a_{i,h})+F_{i,h+1}(\widehat{c}_{i,h+1},\gamma^\star_{i,h+1}(\widehat{c}_{i,h+1}))\right]\\
&=\sum_{i=1}^n\sum_{s_{i,h},p_{i,h},a_{i,h},s_{i,h+1},o_{i,h+1}}\mathbb{P}_{i,h}^{\mathcal{M},c}(s_{i,h},p_{i,h}\,|\,\widehat{c}_{i,h})\gamma_{i,h}(a_{i,h}\,|\,p_{i,h})\mathbb{T}_h(s_{i,h+1}\,|\,s_{i,h},a_{i,h})\\
&\qquad\qquad\mathbb{O}_{i,h+1}(o_{i,h+1}\,|\,s_{i,h+1})\left[r_{i,h}(s_{i,h},a_{i,h})+F_{i,h+1}(\widehat{c}_{i,h+1},\gamma^\star_{i,h+1}(\widehat{c}_{i,h+1}))\right]\\
&:=\sum_{i=1}^n F_{i,h}(\widehat{c}_{i,h},\gamma_{i,h}),
\end{aligned}$$

where for the first equality, we defined $\gamma^\star_{i,h+1}(\widehat{c}_{i,h+1})\in\arg\max_{\gamma_{i,h+1}\in\Gamma_{i,h+1}}F_{i,h+1}(\widehat{c}_{i,h+1},\gamma_{i,h+1})$, thus proving the decomposition. Therefore, to solve Equation (5.1), it suffices to optimize each $F_{i,h}(\widehat{c}_{i,h},\gamma_{i,h})$ individually w.r.t. $\gamma_{i,h}$, which is a linear program with the concatenation of simplex as the constraint by Proposition 8. Thus, the time complexity is $\sum_{i=1}^n\mathrm{poly}(S_i,A_i,P_{i,h})$. $\square$

Remark 3. *In fact, under **Condition 3**, Algorithm 3 and its time complexity can be further improved, where for each $h\in[H]$, we do not necessarily need to enumerate all possible joint approximate common information $\widehat{c}_h$, but only the individual approximate common information $\widehat{c}_{i,h}$ for each $i\in[n]$. This allows the final time complexity to depend only on $\max_{h\in[H]}\sum_{i\in[n]}\widehat{C}_{i,h}P_{i,h}$ instead of $\max_{h\in[H]}\widehat{C}_hP_h$, thus not suffering from the exponential dependency on the number of agents anymore.*

Proof of Theorem 6. The first step is to show that $\widehat{\pi}^\star$, i.e., the return of Algorithm 3 with the equilibrium-computation subroutine replaced as Equation (5.1) is a near-optimal policy for the underlying Dec-POMDP $\mathcal{G}$. To begin with, for any policy $\pi\in\Pi$, we shall prove inductively that for any $h\in[H]$, $c_h\in\mathcal{C}_h$ that

$$V_h^{\pi,\mathcal{M}}(c_h)\le V_h^{\widehat{\pi}^\star,\mathcal{M}}(\widehat{c}_h).$$

It is direct to verify that the inequality holds for $h=H+1$. Now suppose it holds for step $h+1$. For step h, note that

$$\begin{aligned}
V_h^{\pi,\mathcal{M}}(c_h) &= \mathbb{E}_{\{\omega_{j,h}\}_{j\in[n]}}\mathbb{E}^{\mathcal{M}}[\widehat{r}_h^{\mathcal{M}} + V_{h+1}^{\pi,\mathcal{M}}(c_{h+1})\,|\,\widehat{c}_h, \{\pi_{j,h}(\cdot\,|\,\omega_{j,h}, c_h, \cdot)\}_{j\in[n]}] \\
&\le \mathbb{E}_{\{\omega_{j,h}\}_{j\in[n]}}\mathbb{E}^{\mathcal{M}}[\widehat{r}_h^{\mathcal{M}} + V_{h+1}^{\widehat{\pi}^\star,\mathcal{M}}(\widehat{c}_{h+1})\,|\,\widehat{c}_h, \{\pi_{j,h}(\cdot\,|\,\omega_{j,h}, c_h, \cdot)\}_{j\in[n]}] \\
&\le \mathbb{E}^{\mathcal{M}}[\widehat{r}_h^{\mathcal{M}} + V_{h+1}^{\widehat{\pi}^\star,\mathcal{M}}(\widehat{c}_{h+1})\,|\,\widehat{c}_h, \{\widehat{\pi}_{j,h}^\star(\cdot\,|\,\widehat{c}_h, \cdot)\}_{j\in[n]}] \\
&= V_h^{\pi^\star,\mathcal{M}}(\widehat{c}_h),
\end{aligned}$$

where the first inequality is by inductive hypothesis, and the second inequality is due to $V_{h+1}^{\widehat{\pi}^\star,\mathcal{M}}(\widehat{c}_{h+1}) = V_{h+1}^{\star,\mathcal{M}}(\widehat{c}_{h+1})$ and $\{\widehat{\pi}_{j,h}^\star(\cdot\,|\,\widehat{c}_h, \cdot)\}_{j\in[n]}$ is a solution of Equation (5.1). Now under the ground-truth model $\mathcal{G}$, for any $\pi \in \Pi$, $h \in [H]$, $c_h \in \mathcal{C}_h$, by Lemma 2, it holds that

$$V_1^{\pi,\mathcal{G}}(\emptyset) - V_1^{\widehat{\pi}^\star,\mathcal{G}}(\emptyset) \le V_1^{\pi,\mathcal{M}}(\emptyset) - V_1^{\widehat{\pi}^\star,\mathcal{M}}(\emptyset) + 2(H\epsilon_r + H^2\epsilon_z) \le 2(H\epsilon_r + H^2\epsilon_z).$$

To analyze the time complexity, we observe that Algorithm 3 needs to solve Equation (5.1) for $\widehat{C}_h$ times for each $h \in [H]$. Therefore, if Equation (5.1) can be solved with time complexity $\texttt{poly}(S, A, P_h)$ for each $h \in [H]$, the total time complexity of Algorithm 3 is $H \max_{h\in[H]} \widehat{C}_h \times \texttt{poly}(S, A, P_h)$.

Now we are ready to instantiate the guarantees for the examples in Appendix B. Specifically, it is direct to verify that Example 2 and Example 5 together with the approximate belief constructed in Appendix E.4 satisfy **Condition 1** (turned-based structures), while Example 3 and Example 5 together with the approximate belief constructed in Appendix E.4 satisfy **Condition** 2 (the nested information-sharing structure). Therefore, by Proposition 11 and Proposition 12, Equation (5.1) can be solved with time complexity $\texttt{poly}(S, A, P_h)$ for each $h \in [H]$, and the total time complexity of planning such a $2(H\epsilon_r + H^2\epsilon_z)$–team-optimal solution for the Dec-POMDP is $\max_{h\in[H]} \widehat{C}_h \cdot \texttt{poly}(S, A, P_h, H)$. Finally, by Theorem 7, for all examples in Appendix B, there exists an approximate model $\mathcal{M}$ such that $\max\{\epsilon_r, \epsilon_z\} \le \mathcal{O}(\frac{\epsilon}{H^2})$, while $\max_h \widehat{C}_h P_h$ is only quasi-polynomial of the problem instance size. Hence, the time complexity for planning the ϵ-team-optimal solution for those examples is also quasi-polynomial.

For Example 1, i.e., the one-step delayed sharing case, if we additionally assume the **Part (1)** of **Condition 3** (factorized structures) holds, the approximate belief we constructed in Appendix E.4 also satisfies the **Part (2)** of **Condition** 2. Thus, by the improved algorithm and guarantees in Remark 3 and Proposition 13, the total time complexity is $n \max_{i\in[n],h\in[H]} \widehat{C}_{i,h} \times \texttt{poly}(S_i, A_i, P_{i,h}, H)$. Meanwhile, by our construction of the approximate belief, we can ensure $\max\{\epsilon_r, \epsilon_z\} \le \mathcal{O}(\frac{\epsilon}{H^2})$, while $\widehat{C}_{i,h} P_{i,h} \le (A_i O_i)^{\mathcal{O}(\log(SH/\epsilon)/\gamma^4)}$. Therefore, the total time complexity of planning the ϵ-team-optimal solution is $n(A_i O_i)^{\mathcal{O}(\log(SH/\epsilon)/\gamma^4)}$, without suffering from the exponential dependency on n. □

(Quasi-)Efficient learning in Dec-POMDPs without model knowledge. Based on such planning algorithms, we are ready to extend our MARL algorithm to the Dec-POMDP setting for finding the team optimum. Specifically, we only need to replace line 4 of Algorithm 9, i.e., planning for equilibria of the POSG with the planning algorithm for the team-optimal solution of the Dec-POMDP discussed above. Meanwhile, the line 7 of Algorithm 9 for policy selection (Algorithm 7) can be greatly simplified, where we can directly choose

$$\widehat{j} \leftarrow \arg\max_{j\in[K]} R^j,$$

i.e., the policy with the highest empirical rewards. For completeness, we provided the modified algorithm in Algorithm 8. Once the planning algorithm for the approximate model is computationally (quasi-)efficient, the entire learning algorithm (Algorithm 9) is also computationally (quasi-)efficient.

Meanwhile, the procedures of learning the approximate models, i.e., $\{\widehat{\mathcal{M}}(\pi^{1:H,j})\}_{j\in[K]}$ in Algorithm 9 remains unchanged. Therefore, the sample complexity remains the same as learning the equilibria. Formally, we provide the following theorem.

Theorem 21. *Fix $\epsilon, \delta > 0$. Under Assumption 2, for the one-step delayed sharing under the assumption of **Part (1)** in **Condition 3** and all the other information-sharing structures in Appendix B, there exists a multi-agent RL algorithm that learns an ϵ-team optimal solution with probability at least $1-\delta$, in* both *quasi-polynomial time and sample complexities.*

To prove Theorem 21, the major step is to prove the correctness of the simplified policy selection procedure, i.e., the counterpart of Lemma 17 for the Dec-POMDP setting.

Lemma 18. *Fix ϵ, $\delta_3 > 0$. For Algorithm 7, suppose that the K groups of policies $\{\pi^{1:H,j}\}_{j=1}^K$ satisfy that there exists some $j^\star \in [K]$ such that for any policy $\pi \in \Pi$, we have*

$$\left| V_1^{\pi,\mathcal{G}}(\emptyset) - V_1^{\pi,\widehat{\mathcal{M}}(\pi^{1:H,j^\star})}(\emptyset) \right| \leq \epsilon.$$

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 $1-\delta_3$, it holds that

$$V_1^{\pi^{\star,\widehat{j}},\mathcal{G}}(\emptyset) \geq \max_{\pi\in\Pi} V_1^{\pi,\mathcal{G}}(\emptyset) - 4\epsilon.$$

Proof. By the concentration bound on the accumulated rewards of policies $\pi^{\star,j}$, and further a union bound over all $j \in [n]$, with probability at least $1-\delta_3$, the following event $\mathcal{E}_3$ holds for any $j \in [K]$:

$$\left| R^j - V_1^{\pi^{\star,j},\mathcal{G}}(\emptyset) \right| \leq \epsilon.$$

Therefore, it holds that

$$V_1^{\pi^{\star,\widehat{j}},\mathcal{G}}(\emptyset) \geq R^{\widehat{j}} - \epsilon \geq R^{j^\star} - \epsilon \geq V_1^{\pi^{\star,j^\star},\mathcal{G}}(\emptyset) - 2\epsilon.$$

Meanwhile, by denoting $\pi^\star \in \arg\max_{\pi\in\Pi} V_1^{\pi,\mathcal{G}}(\emptyset)$, we have

$$V_1^{\pi^{\star,j^\star},\mathcal{G}}(\emptyset) - V_1^{\pi^\star,\mathcal{G}}(\emptyset) \geq V_1^{\pi^{\star,j^\star},\widehat{\mathcal{M}}(\pi^{1:H,j^\star})}(\emptyset) - V_1^{\pi^\star,\widehat{\mathcal{M}}(\pi^{1:H,j^\star})}(\emptyset) - 2\epsilon \geq -2\epsilon,$$

where the last step is due to the fact that $\pi^{\star,j^\star}$ is the optimal policy of $\widehat{\mathcal{M}}(\pi^{1:H,j^\star})$. Therefore, we conclude that that $V_1^{\pi^{\star,\widehat{j}},\mathcal{G}}(\emptyset) \geq V_1^{\pi^{\star,j^\star},\mathcal{G}}(\emptyset) - 2\epsilon \geq \max_{\pi\in\Pi} V_1^{\pi,\mathcal{G}}(\emptyset) - 4\epsilon$. □

Finally, we are ready to prove Theorem 21.

Proof of Theorem 21. The correctness of the extended learning algorithm follows similarly as the proof of Theorem 8, where for any $\alpha, \delta > 0$, under the exactly the same choices of all parameters (c.f. Appendix E.6) as for learning the equilibrium, with probability $1-\delta$, there exists $j^\star \in [K]$ such that $|V_1^{\pi,\mathcal{G}}(\emptyset) - V_1^{\pi,\widehat{\mathcal{M}}(\pi^{1:H,j^\star})}(\emptyset)| \leq \frac{15\alpha}{200}$ for any $\pi \in \Pi$. Now by Lemma 18, it holds that $V_1^{\pi^{\star,\widehat{j}},\mathcal{G}}(\emptyset) \geq \max_{\pi\in\Pi} V_1^{\pi,\mathcal{G}}(\emptyset) - \frac{60\alpha}{200} \geq \max_{\pi\in\Pi} V_1^{\pi,\mathcal{G}}(\emptyset) - \frac{60\alpha}{200}$, thus concluding that $\pi^{\star,\widehat{j}}$ is an α-team-optimal solution. For the sample complexity, since the choice of all parameters remains the same as that for learning the equilibrium, the sample complexity remains the same as for learning the equilibrium, i.e., quasi-polynomial. For the time complexity, since we have called the planning algorithm, i.e., Algorithm 3 only polynomial times, the total time complexity is also quasi-polynomial by Theorem 6. □