Principled Learning-to-Communicate with Quasi-Classical Information Structures

Xiangyu Liu† Haoyi You† Kaiqing Zhang†

***Abstract*— Learning-to-Communicate (LTC) in partially observable environments has emerged and received increasing attention in deep multi-agent reinforcement learning, where the control and communication strategies are *jointly* learned. On the other hand, the impact of communication has been extensively studied in control theory. In this paper, we seek to formalize and better understand LTC by bridging these two lines of work, through the lens of *information structures* (ISs). To this end, we formalize LTC in decentralized partially observable Markov decision processes (Dec-POMDPs) under the common-information-based framework from decentralized stochastic control, and classify LTC problems based on the ISs before (additional) information sharing. We first show that non-classical LTCs are computationally intractable in general, and thus focus on quasi-classical (QC) LTCs. We then propose a series of conditions for QC LTCs, violating which can cause computational hardness in general. Further, we develop provable planning and learning algorithms for QC LTCs, and show that some examples of QC LTCs satisfying the above conditions can be solved with quasi-polynomial time and samples. Along the way, we also establish some relationship between (strictly) QC IS and the condition of having strategy-independent common-information-based beliefs (SI-CIBs), as well as solving Dec-POMDPs without computationally intractable oracles but beyond those with the SI-CIB condition, which may be of independent interest.**

I. Introduction

The Learning-to-Communicate (LTC) problem has emerged and gained traction in the area of (deep) multi-agent reinforcement learning (MARL) [1], [2], [3]. Unlike classical MARL, which aims to learn *control* strategies that minimize the expected accumulated costs, LTC seeks to *jointly* minimize over both the *control* and the *communication* strategies of all the agents, as a way to mitigate the challenges due to the agents' *partial observability* of the environment. Despite the promising empirical successes, theoretical understandings of LTC remain largely underexplored.

On the other hand, in control theory, a rich literature has investigated the role of *communication* in decentralized/networked control [4], [5], [6], [7], inspiring us to rigorously examine LTCs from such a principled perspective. Most of these studies, however, focused on linear systems, and did not explore the computational or sample complexity guarantees when the system knowledge is not (fully) known. A few recent studies [8], [9] started to explore the settings with general discrete (non-linear) spaces, with special communication protocols and state transition dynamics.

More broadly, the design of communication strategies dictates the *information structure* (IS) of the control system, which characterizes *who knows what and when* [10]. IS and its impact on the *optimization tractability*, especially for linear systems, have been extensively studied in decentralized control, see [11], [12] for comprehensive overviews. In this work, we seek a more principled understanding of LTCs through the lens of information structures, with a focus on the computational and sample complexities of the problem.

Specifically, we formalize LTCs in the general framework of decentralized partially observable Markov decision processes (Dec-POMDPs) [13], as in the empirical works [1], [2], [3]. To achieve finite-time and sample complexity guarantees, we resort to the recent development in [14] on partially observable MARL, based on the common-information-based (CIB) framework [15], [16] from decentralized control, to model the communication and information sharing among agents. We detail our contributions as follows.

Contributions. (i) We formalize Learning-to-Communicate in Dec-POMDPs under the common-information-based framework [15], [16], [14], allowing the sharing of *historical* information, and the modeling of communication costs; (ii) We classify LTCs through the lens of *information structures*, according to the ISs before (additional) information sharing. We then show that LTCs with *non-classical* [11] baseline IS can be computationally intractable. (iii) Given the hardness, we thus focus on *quasi-classical* (QC) LTCs, and propose a series of conditions under which LTCs preserve the QC IS after sharing, while violating which can cause computational hardness in general. (iv) We propose both planning and learning algorithms for QC LTCs, by reformulating them as Dec-POMDPs with *strategy-independent common-information-based beliefs* (SI-CIBs) [16], [14], a condition shown to be critical for tractable computation and learning [14]. (v) Quasi-polynomial time and sample complexities of the algorithms are established for QC LTC examples that satisfy the conditions in (iii). Along the way, we also establish some relationship between *(strictly) quasi-classical* ((s)QC) ISs and the SI-CIB condition in the framework of [16] under certain assumptions, as well as solving general Dec-POMDPs without computationally intractable oracles beyond those with SI-CIBs, and thus advancing the results in [14]. These results may be of independent interest besides studying LTCs. We conclude with some experimental results.

†The authors are ordered alphabetically, and are affiliated with the University of Maryland, College Park, MD, USA, 20742. Emails: {xyliu999, yuriiyou, kaiqing}@umd.edu. This work was supported by the Army Research Office (ARO) grant W911NF-24-1-0085 and the NSF CAREER Award 2443704.

A. Related Work

Communication-control joint optimization. The joint design of control and communication strategies has been studied in the control literature [6], [7], [8], [9]. However, even with model knowledge, the computational complexity (and associated necessary conditions) of solving these models remains elusive, let alone the sample complexity when it comes to learning. Moreover, these models mostly have more special structures, e.g., with linear systems [6], [7], or allowing to share only instantaneous observations [8], [9].

Information sharing and information structures. Information structure has been extensively studied to characterize *who knows what and when* in decentralized control [11], [12]. Our paper aims to formally understand LTC through the lens of information structures. The common-information-based approaches to formalize *information sharing* in [15], [16] serve as the basis of our work. In comparison, these results focused on the *structural results*, without concrete computational (and sample) complexity analysis.

Partially observable MARL theory. Planning and learning in partially observable MARL are known to be hard [17], [18], [19], [13]. Recently, [20], [21] developed polynomial-sample complexity algorithms for partially observable stochastic games, but with computationally intractable oracles; [14] developed quasi-polynomial-time and sample algorithms for such models, leveraging information sharing. In contrast, our paper focuses on *optimizing/learning to share*, together with control strategy optimization/learning.

II. Preliminaries

Notation. We use $\mathbb{N}, \mathbb{Q}, \mathbb{R}$ to denote the sets of all the natural, rational, and real numbers, respectively. For an integer $m > 0$, we denote $[m] := \{1, 2, \cdots, m\}$. For a finite set $\mathcal{X}$, we use $|\mathcal{X}|$ to denote the cardinality of $\mathcal{X}$, and use $\Delta(\mathcal{X})$ to denote the probability simplex over $\mathcal{X}$. For a random variable x, we use $\sigma(x)$ to denote the sigma-algebra generated by x. For σ-algebras $\mathscr{F}_1$ on the space $\mathcal{X}_1$ and $\mathscr{F}_2$ on the space $\mathcal{X}_2$, we denote by $\mathscr{F}_1 \otimes \mathscr{F}_2$ the product σ-algebra on the space $\mathcal{X}_1 \times \mathcal{X}_2$. We use $\mathbb{1}[]$ to denote the indicator function. Unless otherwise noted, the set $\{\}$ considered is ordered, such that elements in the set are indexed.

A. Learning-to-Communicate (with Communication Cost)

For $n > 1$ agents, a (cooperative) *Learning-to-Communicate* problem can be described by a tuple $\mathcal{L} = \langle H, \mathcal{S}, \{\mathcal{A}_{i,h}\}_{i\in[n],h\in[H]}, \{\mathcal{O}_{i,h}\}_{i\in[n],h\in[H]}, \{\mathcal{M}_{i,h}\}_{i\in[n],h\in[H]}, \mathbb{T}, \mathbb{O}, \mu_1, \{\mathcal{R}_h\}_{h\in[H]}, \{\mathcal{K}_h\}_{h\in[H]}\rangle$, where H denotes the length of each episode, and other components are introduced as follows.

1) Decision-making components: We use $\mathcal{S}$ to denote the state space, and $\mathcal{A}_{i,h}$ to denote the *control action* space of agent i at timestep $h \in [H]$. We denote by $s_h \in \mathcal{S}$ the state and by $a_{i,h}$ the control action of agent i at timestep h. We use $a_h := (a_{1,h}, \cdots, a_{n,h}) \in \mathcal{A}_h := \prod_{i\in[n]} \mathcal{A}_{i,h}$ to denote the joint control action of all the n agents at timestep h. We denote by $\mathbb{T} = \{\mathbb{T}_h\}_{h\in[H]}$ the collection of state transition kernels, where $s_{h+1} \sim \mathbb{T}_h(\cdot \mid s_h, a_h) \in \Delta(\mathcal{S})$ at timestep h. We use $\mu_1 \in \Delta(\mathcal{S})$ to denote the initial state distribution. We denote by $\mathcal{O}_{i,h}$ the observation space and by $o_{i,h} \in \mathcal{O}_{i,h}$ the observation of agent i at timestep h. We use $o_h := (o_{1,h}, o_{2,h}, \cdots, o_{n,h}) \in \mathcal{O}_h := \mathcal{O}_{1,h} \times \mathcal{O}_{2,h} \times \cdots \mathcal{O}_{n,h}$ to denote the joint observation of all the n agents at timestep h. We use $\mathbb{O} = \{\mathbb{O}_h\}_{h\in[H]}$ to denote the collection of emission functions, where $o_h \sim \mathbb{O}_h(\cdot \mid s_h) \in \Delta(\mathcal{O}_h)$ at timestep h and state $s_h \in \mathcal{S}$. Also, we denote by $\mathbb{O}_{i,h}(\cdot \mid s_h)$ the emission for agent i, the marginal distribution of $o_{i,h}$ given $\mathbb{O}_h(\cdot \mid s_h)$ for all $s_h \in \mathcal{S}$. At each timestep h, agents will receive a common reward $r_h = \mathcal{R}_h(s_h, a_h)$, where $\mathcal{R}_h : \mathcal{S} \times \mathcal{A}_h \to [0, 1]$ denotes the reward function shared by the agents.

2) Communication components: In addition to reward-driven decision-making, agents also need to decide and learn (what) to communicate with others. At timestep h, agents share part of their information $z_h \in \mathcal{Z}_h$ with other agents, where $\mathcal{Z}_h$ denotes the collection of all possible shared information at timestep h. Here we consider a general setting where the shared information z_h may contain two parts, the *baseline-sharing* part z_h^b that comes from some existing sharing protocol among agents, and the *additional-sharing* part $z_{i,h}^a$ for each agent i that comes from explicit communication *to be decided/learned*, with the joint additional-sharing information $z_h^a := \cup_{i=1}^n z_{i,h}^a$. This general setting covers those considered in most empirical works on LTC [1], [2], [3], with a void baseline sharing part. We kept the baseline sharing since our focus is on the *finite-time* and *sample* tractability of LTC, for which a certain amount of information sharing is known to be necessary [14]. Note that $z_h = z_h^b \cup z_h^a$ and $z_h^b \cap z_h^a = \emptyset$. The shared information is part of the historical observations and (both *control* and *communication*) actions. We denote by $\mathcal{Z}_h^b$, $\mathcal{Z}_h^a$, and $\mathcal{Z}_{i,h}^a$ the collections of all possible z_h^b, z_h^a, and $z_{i,h}^a$ at timestep h.

At timestep h, the *common information* among all the agents is thus defined as the union of all the *shared information* so far: $c_{h^-} = \cup_{t=1}^{h-1} z_t \cup z_h^b$, and $c_{h^+} = \cup_{t=1}^{h} z_t$, where c_{h^-} and c_{h^+} denote the (accumulated) common information *before* and *after* additional sharing, respectively. The *private information* of agent i at time h *before* and *after* additional sharing are denoted by p_{i,h^-}, p_{i,h^+}, respectively, where $p_{i,h^-} \subseteq \{o_{i,1}, a_{i,1}, \cdots, a_{i,h-1}, o_{i,h}\} \backslash c_{h^-}, p_{i,h^+} \subseteq \{o_{i,1}, a_{i,1} \cdots, a_{i,h-1}, o_{i,h}\} \backslash c_{h^+}$. We denote by $p_{h^-} := (p_{1,h^-}, \cdots, p_{n,h^-})$ the joint private information *before* additional sharing, by $p_{h^+} := (p_{1,h^+}, \cdots, p_{n,h^+})$ the joint private information *after* additional sharing, at timestep h. We then denote by $\tau_{i,h^-} := p_{i,h^-} \cup c_{h^-}, \tau_{i,h^+} := p_{i,h^+} \cup c_{h^+}$ the *information available* to agent i at timestep h, before and after additional sharing, respectively, with $\tau_{h^-} := p_{h^-} \cup c_{h^-}, \tau_{h^+} := p_{h^+} \cup c_{h^+}$ denoting the associated joint information. We use $\mathcal{C}_{h^-}, \mathcal{C}_{h^+}, \mathcal{P}_{i,h^-}, \mathcal{P}_{i,h^+}, \mathcal{P}_{h^-}, \mathcal{P}_{h^+}, \mathcal{T}_{i,h^-}, \mathcal{T}_{i,h^+}, \mathcal{T}_{h^-}, \mathcal{T}_{h^+}$ to denote, respectively, the corresponding collections of all possible $c_{h^-}, c_{h^+}, p_{i,h^-}, p_{i,h^+}, p_{h^-}, p_{h^+}, \tau_{i,h^-}, \tau_{i,h^+}, \tau_{h^-}, \tau_{h^+}$.

We use $m_{i,h}$ to denote the *communication action* of agent i at timestep h, and it will determine what information $z_{i,h}^a$ she will share, through the way specified later. We denote by $\mathcal{M}_{i,h}$ the space of $m_{i,h}$, and by $m_h := (m_{1,h}, \cdots, m_{n,h}) \in$

$\mathcal{M}_h := \mathcal{M}_{1,h} \times \cdots \mathcal{M}_{n,h}$ the joint communication action of all the agents. $\mathcal{K}_h : \mathcal{Z}_h^a \to [0,1]$ denotes the *communication cost* function, and $\kappa_h = \mathcal{K}_h(z_h^a)$ denotes the incurred communication cost at timestep h, due to additional sharing.

3) System evolution: The system evolves by alternating between the communication and the control steps as follows.

Communication step: At each timestep h, each agent i observes $o_{i,h}$ and may share part of her private information via baseline sharing, receives the baseline sharing of information from others, and forms p_{i,h^-} and c_{h^-}. Then, each agent i chooses her communication action, which determines the additional sharing of information, receives the additional-sharing of information from others, forms p_{i,h^+} and c_{h^+}, and incurs some communication cost κ_h. Formally, the evolution of the information is formalized as follows, which, unless otherwise noted, will be assumed throughout the paper. We follow the convention that any quantity at $h = 0$ is empty/null.

Assumption II.1 (*Information evolution*)**.** For each $h \in [H]$,

(a) (Baseline sharing). $z_h^b = \chi_h(p_{(h-1)^+}, a_{h-1}, o_h)$ for some fixed transformation χ_h;

(b) (Additional sharing). For each agent $i \in [n], z_{i,h}^a = \phi_{i,h}(p_{i,h^-}, m_{i,h})$ for some function $\phi_{i,h}$, given communication action $m_{i,h}$, and $m_{i,h} \in z_{i,h}^a$; and the joint sharing $z_h^a := \cup_{i\in[n]} z_{i,h}^a$ is thus generated by $z_h^a = \phi_h(p_{h^-}, m_h)$, for some function ϕ_h;

(c) (Private information before sharing). For each agent $i \in [n]$, $p_{i,h^-} = \xi_{i,h}(p_{i,(h-1)^+}, a_{i,h-1}, o_{i,h})$ for some fixed transformation $\xi_{i,h}$, and the joint private information thus evolves as $p_{h^-} = \xi_h(p_{(h-1)^+}, a_{h-1}, o_h)$ for some fixed transformation ξ_h;

(d) (Private information after sharing). For each agent $i \in [n]$, $p_{i,h^+} = p_{i,h^-} \backslash z_{i,h}^a$;

(e) ($(\tau_{i,h^-}, \tau_{i,h^+})$-inclusion). For each agent $i \in [n]$, $\tau_{i,h^-} \subseteq \tau_{i,h^+} \subseteq \tau_{i,(h+1)^-}$, and $o_{i,h} \in \tau_{i,h^-}$.

Note that as *fixed transformations* (e.g., χ_h and $\xi_{i,h}$ above), they are not affected by the *realized values* of the random variables, but dictate some *pre-defined* transformation of the input random variables. See [15], [16] and §B in [14] for common examples of baseline sharing that admit such fixed transformations when there is no additional sharing, and examples in §A on how they are extended in the LTC setting. It should not be confused with some general *function* (e.g., $\phi_{i,h}$ above), which may depend on the *realized values* of the input random variables. (a) and (c) on baseline sharing follow from those in [16], [14]; (b) and (d) on additional sharing dictate how the communication action affects the sharing based on private information. For example, a common choice of $(\mathcal{M}_{i,h}, \phi_{i,h})$ is that $\mathcal{M}_{i,h} = \{0,1\}^{|p_{i,h^-}|}$, for any $p_{i,h^-} \in \mathcal{P}_{i,h^-}$ and $m_{i,h} \in \mathcal{M}_{i,h}$, $\phi_{i,h}(p_{i,h^-}, m_{i,h})$ consists of the k-th element ($k \in [|p_{i,h^-}|]$) of p_{i,h^-} if and only if the k-th element of $m_{i,h}$ is 1. As $m_{i,h}$ (depicting what to share) will be known given $z_{i,h}^a$ (what has been shared), $m_{i,h}$ is thus also modeled as being shared, i.e., $m_{i,h} \in z_{i,h}^a$. This is also consistent with the models in [8], [9] on control/communication joint optimization. (e) means that the agent has full memory of the information she had in the past and at present. We emphasize that this is closely related, but different from the common notion of *perfect recall* [22], where the agent has to recall all her own *past actions*. Condition (e), in contrast, relaxes the memorization of the actions, but includes the instantaneous observation $o_{i,h}$. This condition is satisfied by the models and examples in [11], [15], [16], [14]. See also §A for more examples that satisfy this assumption. Meanwhile, for both the baseline and additional sharing protocols, we follow the model in the series of works on partial history/information sharing [15], [16], [14], [8], [9] that, if an agent shares, she will share the information with *all other* agents, and and make it compatible with information structure literature, where σ-algebra is considered.

Assumption II.2. $\forall i_1, i_2 \in [n], h_1, h_2 \in [H], i_1 \neq i_2, h_1 < h_2$, if $\sigma(o_{i_1,h_1}) \subseteq \sigma(\tau_{i_2,h_2^-})$, then $\sigma(o_{i_1,h_1}) \subseteq \sigma(c_{h_2^-})$, and if $\sigma(a_{i_1,h_1}) \subseteq \sigma(\tau_{i_2,h_2^-})$, then $\sigma(a_{i_1,h_1}) \subseteq \sigma(c_{h_2^-})$; if $\sigma(o_{i_1,h_1}) \subseteq \sigma(\tau_{i_2,h_2^+})$, then $\sigma(o_{i_1,h_1}) \subseteq \sigma(c_{h_2^+})$, and if $\sigma(a_{i_1,h_1}) \subseteq \sigma(\tau_{i_2,h_2^+})$, then $\sigma(a_{i_1,h_1}) \subseteq \sigma(c_{h_2^+})$.

Assumptions II.1-II.2 will be made throughout the paper.

Decision-making step: After the communication, each agent i chooses her control action $a_{i,h}$, receives a reward r_h, and the joint action a_h drives the state to $s_{h+1} \sim \mathbb{T}_h(\cdot \mid s_h, a_h)$.

4) Strategies and solution concept: At timestep h, each agent i has two strategies, a *control* strategy and a *communication* strategy. We define a control strategy as $g_{i,h}^a : \mathcal{T}_{i,h^+} \to \mathcal{A}_{i,h}$ and a communication strategy as $g_{i,h}^m : \mathcal{T}_{i,h^-} \to \mathcal{M}_{i,h}$. We denote by $g_h^a = (g_{1,h}^a, \cdots, g_{n,h}^a)$ the joint control strategy and by $g_h^m = (g_{1,h}^m, \cdots, g_{n,h}^m)$ the joint communication strategy. We denote by $\mathcal{G}_{i,h}^a, \mathcal{G}_{i,h}^m, \mathcal{G}_h^a, \mathcal{G}_h^m$ the corresponding spaces of $g_{i,h}^a, g_{i,h}^m, g_h^a, g_h^m$, respectively.

The objective of the agents in the LTC problem is to maximize the expected accumulated sum of the reward and the negative communication cost from timestep $h = 1$ to H:

$$J_{\mathcal{L}}(g_{1:H}^a, g_{1:H}^m) := \mathbb{E}_{\mathcal{L}}\left[\sum_{h=1}^{H}(r_h - \kappa_h) \,\middle|\, g_{1:H}^a, g_{1:H}^m\right],$$

where the expectation $\mathbb{E}_{\mathcal{L}}$ is taken over all the randomness in the system evolution, given the strategies $(g_{1:H}^a, g_{1:H}^m)$. With this objective, for any $\epsilon \geq 0$, we can define the solution concept of an *ϵ-team optimum* for $\mathcal{L}$ as follows.

Definition II.3 (ϵ-team optimum)**.** We call a joint strategy $(g_{1:H}^a, g_{1:H}^m)$ an ϵ-team optimal strategy of the LTC $\mathcal{L}$ if

$$\max_{\widetilde{g}_{1:H}^a \in \mathcal{G}_{1:H}^a, \widetilde{g}_{1:H}^m \in \mathcal{G}_{1:H}^m} J_{\mathcal{L}}(\widetilde{g}_{1:H}^a, \widetilde{g}_{1:H}^m) - J_{\mathcal{L}}(g_{1:H}^a, g_{1:H}^m) \leq \epsilon.$$

B. Information Structures of LTC

In decentralized stochastic control, the notion of information structure [23], [11] captures *who knows what and when* as the system evolves. In LTC, as the additional sharing via communication will also affect the IS and is *not* determined *beforehand*, when we discuss the *IS of an LTC problem*, we will refer to that of the problem *with only baseline sharing*. In

particular, an LTC $\mathcal{L}$ without additional sharing is essentially a Dec-POMDP (with potential baseline information sharing), as defined in §E for completeness. We formally define such a Dec-POMDP *induced* by $\mathcal{L}$ as follows.

Definition II.4 (Dec-POMDP (with information sharing) induced by LTC)**.** For an LTC $\mathcal{L} = \langle H, \mathcal{S}, \{\mathcal{A}_{i,h}\}_{i\in[n],h\in[H]}, \{\mathcal{O}_{i,h}\}_{i\in[n],h\in[H]}, \{\mathcal{M}_i\}_{i\in[n]}, \mathbb{T}, \mathbb{O}, \mu_1, \{\mathcal{R}_h\}_{h\in[H]}, \{\mathcal{K}_h\}_{h\in[H]}\rangle$, we call a Dec-POMDP (with information sharing) $\overline{\mathcal{D}}_{\mathcal{L}}$ *the Dec-POMDP (with information sharing) induced by* $\mathcal{L}$ if the agents share information only following the baseline sharing protocol of $\mathcal{L}$, i.e., without additional sharing, which can be characterized by the tuple $\overline{\mathcal{D}}_{\mathcal{L}} := \langle H, \mathcal{S}, \{\mathcal{A}_{i,h}\}_{i\in[n],h\in[H]}, \{\mathcal{O}_{i,h}\}_{i\in[n],h\in[H]}, \mathbb{T}, \mathbb{O}, \mu_1, \{\mathcal{R}_h\}_{h\in[H]}\rangle$. We may refer to it as the *Dec-POMDP induced by LTC* or the *induced Dec-POMDP* for short.

In §II-A, we introduced LTC in the *state-space model*. Information structure is oftentimes more conveniently discussed under the equivalent framework of *intrinsic models* [23] (see the instantiation for Dec-POMDPs in §F for completeness). In an intrinsic model, each agent only *acts once* throughout the system evolution, and the same agent in the state-space model at different timesteps is now treated as *different agents*. There are thus $n \times H$ agents in total. Formally, for completeness, we extend the intrinsic-model-based reformulation to LTCs in §F.

(Strictly) quasi-classical ISs are important subclasses of ISs, which were first introduced for decentralized stochastic control [23], [24], [12] (see the instantiation for Dec-POMDPs in §F). An IS that is not QC is *non-classical* [11], [12]. We extend such a categorization to LTC problems with different ISs as follows.

Definition II.5 ((Strictly) quasi-classical LTC)**.** We call an LTC $\mathcal{L}$ *(strictly) quasi-classical* if the Dec-POMDP induced by $\mathcal{L}$ (see Definition II.4) is *(strictly) quasi-classical.* Namely, each agent in the intrinsic model of $\overline{\mathcal{D}}_{\mathcal{L}}$ knows the information (and the actions) of the agents who influence her, either directly or indirectly.

Similarly, an LTC $\mathcal{L}$ that is not QC is called *non-classical*. See §A for examples of QC and sQC LTC. Note that the categorization above is defined based on the ISs *before* additional sharing, as an inherent property of the LTC problem, since additional sharing is the solution *to be* decided/learned. We focus on finding such a solution next.

III. Hardness and Structural Assumptions

It is known that computing an (approximate) team-optimum in Dec-POMDPs, which are LTCs *without* information-sharing, is `NEXP-hard` [13]. The hardness cannot be fully circumvented even when agents are allowed to share information: even if agents share all the information, the LTC problem becomes a Partially Observable Markov Decision Process (POMDP), which is known to be `PSPACE-hard` [17], [18]. Hence, additional assumptions are necessary to make LTCs computationally tractable. We introduce several such assumptions and their justifications below, whose proofs can be found in §B.

Recently, [25] showed that *observable* POMDPs [26], a class of POMDPs with relatively *informative* observations, admit *quasi-polynomial time* algorithms to solve. Such a condition and quasi-polynomial complexity result was then established for Dec-POMDPs with information sharing in [14]. As solving LTCs is at least as hard as solving the Dec-POMDPs considered in [14], we first also make such an observability assumption on the *joint* emission function as in [14], to avoid computationally intractable oracles.

Assumption III.1 (γ-observability [26], [25], [14])**.** There exists a $\gamma > 0$ such that $\forall h \in [H]$, the emission $\mathbb{O}_h$ satisfies that $\forall b_1, b_2 \in \Delta(\mathcal{S})$, $\left\|\mathbb{O}_h^\top b_1 - \mathbb{O}_h^\top b_2\right\|_1 \geq \gamma \left\|b_1 - b_2\right\|_1$.

However, we show next that, Assumption III.1 is not enough when it comes to LTC, if the baseline sharing IS is not favorable, in particular, *non-classical* [11]. The hardness persists even under a few additional assumptions to be introduced later that will make LTCs tractable.

Lemma III.2 (Non-classical LTCs are hard)**.** For non-classical LTCs under Assumption III.1, III.4, III.5, and III.7, finding an $\frac{\epsilon}{H}$-team optimum is `PSPACE-hard`.

Note that the hardness comes from the intuition that, when communication costs are high, the additional sharing from LTC will be limited, preventing the upgrade of the IS from a non-classical one to a (quasi-)classical one, which is hard with only the *joint* observability of the emission (see Assumption III.1), even along with several other assumptions.

By Lemma III.2, we will hence focus on the *quasi-classical* LTCs hereafter. Indeed, QC is also known to be critical for efficiently solving *continuous-space* and *linear* decentralized control [27], [28]. However, quasi-classicality may not be sufficient for LTCs, since the additional sharing may *break* the QC IS, and introduce computational hardness, as argued below.

Firstly, the breaking may result from the *communication strategies*. In particular, the general communication strategy space in §II-A.4 allows the dependence on agents' *private information*, which introduces incentives for *signaling* [11] and can also cause computational hardness, as shown next.

Lemma III.3 (QC LTCs with full-history-dependent communication strategies are hard)**.** For QC LTCs under Assumption III.1, together with Assumptions III.5, and III.7, computing a team-optimum in the general space of $(\mathcal{G}^a_{1:H}, \mathcal{G}^m_{1:H})$ with $\mathcal{G}^m_{i,h} := \{g^m_{i,h} : \mathcal{T}_{i,h^-} \to \mathcal{M}_{i,h}\}$ is `NP-hard`.

The hardness in Lemma III.3 originates from the fact that when depending on the private/local information, determining the communication action can be made as a *Team Decision problem* (TDP) [29], which is known to be hard. This will be the case even when the instantaneous observations are relatively observable (see Assumptions III.1-III.7).

To avoid this hardness, we thus focus on communication strategies that only condition on the *common information*. Intuitively, this assumption is not unreasonable, as it means

that *which historical information to share* is determined by *what has been shared* (in the common information). Note that, this does not lose the generality in the sense that the private information p_{i,h^-} *can still be shared*. It only means that the communication action is not determined based on p_{i,h^-}, and the additional sharing is still dictated by $z^a_{i,h} = \phi_{i,h}(p_{i,h^-}, m_{i,h})$ (see Assumption II.1), depending on p_{i,h^-}.

Assumption III.4 (Common-information-based communication strategy)**.** The communication strategies take *common information* as input, with the following form:

$$\forall i \in [n], h \in [H], \quad g^m_{i,h} : \mathcal{C}_{h^-} \to \mathcal{M}_{i,h}. \tag{III.1}$$

Secondly, the breaking of QC may result from the *control strategies*: if some agent did *not* influence others in the baseline sharing (and thus these other agents did *not* have to access the agent's available information, while still satisfying QC), while she starts to influence others by *sharing* her (*useless*) *control* actions, this will make her *control strategies* relevant. We make the following two assumptions to avoid the related pessimistic cases, each followed by a computational hardness result when the condition is missing.

Specifically, in some special cases, the action of some agents may not influence the state transition. Such actions are thus *useless* in terms of decision-making, when there is *no* information sharing. However, if they were deemed *non-influential*, but shared via additional sharing, then QC may break for the LTC problem. We thus make the following assumption, followed by a justification result.

Assumption III.5 (Control-useless action is not used)**.** $\forall i \in [n], h \in [H]$, if agent i's action $a_{i,h}$ does not influence the state s_{h+1}, namely, $\forall s_h \in \mathcal{S}, a_h \in \mathcal{A}_h, a'_{i,h} \in \mathcal{A}_{i,h}, a'_{i,h} \neq a_{i,h}, \mathbb{T}_h(\cdot \,|\, s_h, a_h) = \mathbb{T}_h(\cdot \,|\, s_h, (a'_{i,h}, a_{-i,h}))$. Then, $\forall h' > h$, the random variable $a_{i,h} \notin \tau_{h'^-}$ and $a_{i,h} \notin \tau_{h'^+}$.

Lemma III.6 (QC LTCs without Assumption III.5 are hard)**.** For QC LTCs under Assumptions III.1, III.4, and III.7, finding a team-optimum is still `NP-hard`.

Note that other than the justification above based on computational hardness, Assumption III.5 has been *implicitly* made in the IS examples in the literature when there are *uncontrolled* state dynamics, see e.g., [16], [14]. Moreover, we emphasize that for common cases where actions *do* affect the state transition, this assumption becomes not necessary.

Other than *not influencing* state transition, an action may also be non-influential if the emission functions of other agents are *degenerate*: they cannot *sense* the influence from previous agents' actions. We thus make the following assumption on the emissions, followed by a justification result.

Assumption III.7 (Other agents' emissions are non-degenerate)**.** $\forall h \in [H], i \in [n]$, $\mathbb{O}_{-i,h}$ satisfies $\forall b_1, b_2 \in \Delta(\mathcal{S})$ such that $b_1 \neq b_2$, $\mathbb{O}^\top_{-i,h} b_1 \neq \mathbb{O}^\top_{-i,h} b_2$.

Lemma III.8 (QC LTCs without Assumption III.7 are hard)**.** For QC LTCs under Assumption III.1, III.4, and III.5, finding an ϵ/H-team optimum is still `PSPACE-hard`.

We have justified the above assumptions by showing that missing one of them may cause computational intractability of LTCs in general. More importantly, as we will show later, as another justification, LTCs under Assumptions III.4, III.5, and III.7 can indeed *preserve* the QC/sQC information structure *after* additional sharing, making it possible for the overall LTC problem to be computationally tractable. More examples that satisfy these assumptions can also be found in §A.

IV. Solving LTC Problems Provably

We now study how to solve LTC provably, via either *planning* (with model knowledge) or *learning* (without model knowledge). The process of our solution is shown in Fig. 1 and proofs of the results can be found in §C.

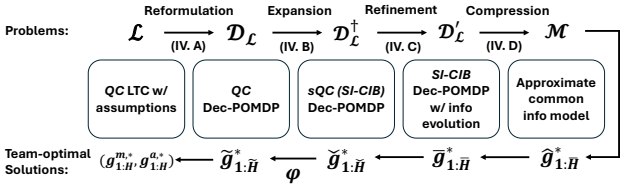


Fig. 1: Illustrating the subroutines 1 for solving the LTC problems.

A. An Equivalent Dec-POMDP

Given any LTC $\mathcal{L}$, we can define a Dec-POMDP $\mathcal{D}_\mathcal{L}$ characterized by $\langle \widetilde{H}, \widetilde{\mathcal{S}}, \{\widetilde{\mathcal{A}}_{i,h}\}_{i\in[n],h\in[\widetilde{H}]}, \{\widetilde{\mathcal{O}}_{i,h}\}_{i\in[n],h\in[\widetilde{H}]}, \{\widetilde{\mathbb{T}}_h\}_{h\in[\widetilde{H}]}, \{\widetilde{\mathbb{O}}_h\}_{h\in[\widetilde{H}]}, \widetilde{\mu}_1, \{\widetilde{\mathcal{R}}_h\}_{h\in[\widetilde{H}]}\rangle$, such that these two are equivalent (under the assumptions in §III): $\forall h \in [H]$,

$$\begin{aligned}
&\widetilde{H} = 2H, \quad \widetilde{\mathcal{S}} = \mathcal{S}, \quad \widetilde{s}_{2h-1} = \widetilde{s}_{2h} = s_h, \quad \widetilde{\mathcal{A}}_{i,2h-1} = \mathcal{M}_{i,h}, \\
&\widetilde{\mathcal{A}}_{i,2h} = \mathcal{A}_{i,h}, \quad \widetilde{\mathcal{O}}_{i,2h-1} = \mathcal{O}_{i,h}, \quad \widetilde{\mathcal{O}}_{i,2h} = \{\emptyset\}, \quad \widetilde{\mu}_1 = \mu_1, \\
&\widetilde{\mathbb{O}}_{2h-1} = \mathbb{O}_h, \quad \widetilde{\mathbb{T}}_{2h-1}(\widetilde{s}_{2h} \,|\, \widetilde{s}_{2h-1}, \widetilde{a}_{2h-1}) = \mathbb{1}[\widetilde{s}_{2h} = \widetilde{s}_{2h-1}], \\
&\widetilde{\mathbb{T}}_{2h}(\widetilde{s}_{2h+1} \,|\, \widetilde{s}_{2h}, \widetilde{a}_{2h}) = \mathbb{T}_h(\widetilde{s}_{2h+1} \,|\, \widetilde{s}_{2h}, \widetilde{a}_{2h}), \\
&\widetilde{\mathcal{R}}_{2h-1} = -\mathcal{K}_h, \quad \widetilde{\mathcal{R}}_{2h} = \mathcal{R}_h, \quad \widetilde{p}_{i,2h-1} = \emptyset, \quad \widetilde{p}_{i,2h} = p_{i,h^+}, \\
&\widetilde{c}_{2h-1} = c_{h^-}, \ \widetilde{c}_{2h} = c_{h^+}, \ \widetilde{z}_{2h-1} = z^b_h, \ \widetilde{z}_{2h} = z^a_h.
\end{aligned} \tag{IV.1}$$

for all $(i,h) \in [n] \times [H]$, $s_h \in \mathcal{S}, a_{i,h} \in \mathcal{A}_{i,h}, o_{i,h} \in \mathcal{O}_{i,h}, m_{i,h} \in \mathcal{M}_{i,h}, p_{i,h^-} \in \mathcal{P}_{i,h^-}, p_{i,h^+} \in \mathcal{P}_{i,h^+}, c_{h^-} \in \mathcal{C}_{h^-}, c_{h^+} \in \mathcal{C}_{h^+}, \tau_{i,h^-} \in \mathcal{T}_{i,h^-}, \tau_{i,h^+} \in \mathcal{T}_{i,h^+}$. Note that we follow the convention of $\widetilde{\tau}_{i,h} := \widetilde{p}_{i,h} \cup \widetilde{c}_h$ for any $h \in [\widetilde{H}]$, and at the odd timestep $2t-1$ for any $t \in [H]$, we have $\widetilde{p}_{i,2t-1} = \emptyset$ under Assumption III.4, i.e., in $\mathcal{D}_\mathcal{L}$, each agent only uses the common information so far for decision-making at timestep $2h-1$. Correspondingly, for any $h \in [\widetilde{H}], i \in [n]$, we denote by $\widetilde{g}_{i,h}, \widetilde{g}_h$ the (joint) strategy and by $\widetilde{\mathcal{G}}_{i,h}, \widetilde{\mathcal{G}}_h$ the (joint) strategy spaces. Similarly, the objective of $\mathcal{D}_\mathcal{L}$ is defined as $J_{\mathcal{D}_\mathcal{L}}(\widetilde{g}_{1:\widetilde{H}}) = \mathbb{E}_{\mathcal{D}_\mathcal{L}}[\sum_{h=1}^{\widetilde{H}} \widetilde{r}_h \,|\, \widetilde{g}_{1:\widetilde{H}}]$.

Essentially, this reformulation splits the H-step decision-making and communication procedure into a $2H$-step one. A similar splitting of the timesteps was also used in [8], [9]. In comparison, we consider a more general setting, where the state is not decoupled, and agents are allowed to share the observations and actions at the *previous* timesteps, due

to the generality of our LTC formulation. The equivalence between $\mathcal{L}$ and $\mathcal{D}_{\mathcal{L}}$ is more formally stated as follows.

Proposition IV.1 (Equivalence between $\mathcal{L}$ and $\mathcal{D}_{\mathcal{L}}$)**.** Let $\mathcal{D}_{\mathcal{L}}$ be the reformulated Dec-POMDP from $\mathcal{L}$ with Assumption III.4, then the solutions of the two problems are equivalent, in the sense that $\forall g^m_{1:H} \in \mathcal{G}^m_{1:H}, g^a_{1:H} \in \mathcal{G}^a_{1:H}, i \in [n]$, let $\widetilde{g}_{1:\widetilde{H}} = (g^m_1, g^a_1, \cdots, g^m_H, g^a_H)$, then $J_{\mathcal{D}_{\mathcal{L}}}(\widetilde{g}_{1:\widetilde{H}}) = J_{\mathcal{L}}(g^m_{1:H}, g^a_{1:H})$. Also, $\forall \widetilde{g}_{1:\widetilde{H}} \in \widetilde{\mathcal{G}}_{1:\widetilde{H}}, i \in [n]$, let $g^m_{1:H} = (\widetilde{g}_1, \widetilde{g}_3, \cdots, \widetilde{g}_{\widetilde{H}-1}), g^a_{1:H} = (\widetilde{g}_2, \widetilde{g}_4, \cdots, \widetilde{g}_{\widetilde{H}})$, then $J_{\mathcal{L}}(g^m_{1:H}, g^a_{1:H}) = J_{\mathcal{D}_{\mathcal{L}}}(\widetilde{g}_{1:\widetilde{H}})$.

Also, the Dec-POMDP $\mathcal{D}_{\mathcal{L}}$ preserves the QC IS from $\mathcal{L}$.

Theorem IV.2 (Preserving (s)QC)**.** If $\mathcal{L}$ is (s)QC and satisfies Assumptions III.4, III.5, and III.7, then the reformulated Dec-POMDP $\mathcal{D}_{\mathcal{L}}$ is also (s)QC.

By Proposition IV.1, it suffices to solve the reformulated $\mathcal{D}_{\mathcal{L}}$ that are QC/sQC, which will be our focus next.

B. Strict Expansion of $\mathcal{D}_{\mathcal{L}}$

However, being QC does not necessarily imply $\mathcal{D}_{\mathcal{L}}$ can be solved *without* computationally intractable oracles. Note that this is different from the continuous-space, linear quadratic case, where QC problems can be reformulated and solved efficiently [27], [28]. With discrete spaces, the recent result [14] established a concrete *quai-polynomial*-time complexity for planning, under the *strategy independence* assumption [16] on the common-information-based beliefs [15], [16]. This SI-CIB assumption was shown critical for *computational tractability* [14] – it eliminates the need to *enumerate* the past strategies in dynamic programming, which would otherwise be prohibitively large. Thus, we need to connect QC IS to the SI-CIB condition for computational tractability.

Interestingly, under certain conditions, one can connect these two conditions for the reformulated Dec-POMDP $\mathcal{D}_{\mathcal{L}}$. As the first step, we will *expand* the QC $\mathcal{D}_{\mathcal{L}}$ by adding the *actions* of the agents who influence the later agents in the intrinsic model of $\mathcal{D}_{\mathcal{L}}$ to the shared information. We denote the strictly expanded Dec-POMDP as $\mathcal{D}^{\dagger}_{\mathcal{L}}$. We replace the $\widetilde{\ }$ notation in $\mathcal{D}_{\mathcal{L}}$ by the $\breve{\ }$ notation in $\mathcal{D}^{\dagger}_{\mathcal{L}}$. The horizon, states, actions, observations, transitions, and reward functions remain the same, but the sets of information $\breve{p}_h, \breve{c}_h, \breve{\tau}_h, \breve{p}_{i,h}, \breve{\tau}_{i,h}$ are different: for any $h \in [\widetilde{H}], i \in [n]$

$$\begin{aligned} \breve{c}_h &= \widetilde{c}_h \cup \{\widetilde{a}_{j,t} \,|\, j \in [n], t < h, \sigma(\widetilde{\tau}_{j,t}) \subseteq \sigma(\widetilde{c}_h)\} \\ \breve{p}_{i,h} &= \widetilde{p}_{i,h} \backslash \{\widetilde{a}_{i,t} \,|\, t < h, \sigma(\widetilde{\tau}_{i,t}) \subseteq \sigma(\widetilde{c}_h)\}, \end{aligned} \tag{IV.2}$$

and we follow the convention to define $\breve{\tau}_{i,h} := \breve{p}_{i,h} \cup \breve{c}_h$. It is not hard to verify the following.

Lemma IV.3. If $\mathcal{D}_{\mathcal{L}}$ is QC, then $\mathcal{D}^{\dagger}_{\mathcal{L}}$ is sQC.

In contrast to the reformulation in §IV-A, the expansion here cannot guarantee the equivalence between $\mathcal{D}_{\mathcal{L}}$ and $\mathcal{D}^{\dagger}_{\mathcal{L}}$: the strategy spaces of $\mathcal{D}^{\dagger}_{\mathcal{L}}$ are larger than those of $\mathcal{D}_{\mathcal{L}}$, as each agent can now access more information, i.e., $\widetilde{\tau}_{i,h} \subseteq \breve{\tau}_{i,h}$. Fortunately, the team-optimal value and strategy of both Dec-POMDPs are related, as shown in the following theorem.

Theorem IV.4. Let $\mathcal{D}_{\mathcal{L}}$ be the QC Dec-POMDP reformulated from a QC LTC $\mathcal{L}$, and $\mathcal{D}^{\dagger}_{\mathcal{L}}$ be the sQC expansion of $\mathcal{D}_{\mathcal{L}}$. Then, for any ϵ-team-optimal strategy $\breve{g}^*_{1:\breve{H}}$ of $\mathcal{D}^{\dagger}_{\mathcal{L}}$, there exists a function φ such that $\widetilde{g}^*_{1:\widetilde{H}} = \varphi(\breve{g}^*_{1:\breve{H}}, \mathcal{D}_{\mathcal{L}})$ is an ϵ-team-optimal strategy of $\mathcal{D}_{\mathcal{L}}$, with $J_{\mathcal{D}_{\mathcal{L}}}(\widetilde{g}^*_{1:\widetilde{H}}) = J_{\mathcal{D}^{\dagger}_{\mathcal{L}}}(\breve{g}^*_{1:\breve{H}})$.

Theorem IV.4 shows that one can solve the QC $\mathcal{D}_{\mathcal{L}}$ by first solving the sQC expansion $\mathcal{D}^{\dagger}_{\mathcal{L}}$, and then using an oracle φ to translate it back as a solution in the strategy spaces of $\mathcal{D}_{\mathcal{L}}$, without loss of optimality. Importantly, we show in Algorithm 4 that how to implement such a φ function efficiently.

As shown below, a benefit of obtaining an *sQC* $\mathcal{D}^{\dagger}_{\mathcal{L}}$ is that, it also has *SI-CIBs*, making it possible to be solved without computationally intractable oracles as in [14].

Theorem IV.5. Let $\mathcal{D}^{\dagger}_{\mathcal{L}}$ be an sQC Dec-POMDP generated from $\mathcal{L}$ that satisfies Assumptions III.4, III.5 and III.7 after reformulation and strict expansion, then $\mathcal{D}^{\dagger}_{\mathcal{L}}$ has *strategy-independent common-information-based beliefs* [16], [14]. More formally, for any $h \in [\breve{H}]$, any two different joint strategies $\breve{g}_{1:h-1}$ and $\breve{g}'_{1:h-1}$, and any common information $\breve{c}_h$ that can be reached under both $\breve{g}_{1:h-1}$ and $\breve{g}'_{1:h-1}$, for any joint private information $\breve{p}_h \in \breve{\mathcal{P}}_h$ and state $\breve{s}_h \in \breve{\mathcal{S}}$,

$$\mathbb{P}^{\mathcal{D}^{\dagger}_{\mathcal{L}}}_h(\breve{s}_h, \breve{p}_h \,|\, \breve{c}_h, \breve{g}_{1:h-1}) = \mathbb{P}^{\mathcal{D}^{\dagger}_{\mathcal{L}}}_h(\breve{s}_h, \breve{p}_h \,|\, \breve{c}_h, \breve{g}'_{1:h-1}). \tag{IV.3}$$

C. Refinement of $\mathcal{D}^{\dagger}_{\mathcal{L}}$

Despite having SI-CIBs, $\mathcal{D}^{\dagger}_{\mathcal{L}}$ is still not eligible for applying the results in [14]: the information evolution rules of $\mathcal{D}^{\dagger}_{\mathcal{L}}$ break those in [16], [14]. Specifically, due to Assumption III.4, we set $\widetilde{\tau}_{i,2t-1} = \widetilde{c}_{2t-1}, \widetilde{p}_{i,2t-1} = \emptyset, \forall t \in [H], i \in [n]$ in $\mathcal{D}_{\mathcal{L}}$, which violates Assumption 1 in [16], [14]. To address this issue, we propose to further *refine* the $\mathcal{D}^{\dagger}_{\mathcal{L}}$ to obtain a Dec-POMDP $\mathcal{D}'_{\mathcal{L}}$, which satisfies the information evolution rules. We replace the $\breve{\ }$ notation in $\mathcal{D}^{\dagger}_{\mathcal{L}}$ by the $\bar{\ }$ notation in $\mathcal{D}'_{\mathcal{L}}$. The elements in $\mathcal{D}'_{\mathcal{L}}$ remain the same as those in $\mathcal{D}^{\dagger}_{\mathcal{L}}$, except that the private information at odd steps is now refined as: for any $t \in [H]$

$$\overline{p}_{i,2t-1} = p_{i,t^-} \backslash \breve{c}_{2t-1}, \tag{IV.4}$$

and we define $\overline{\tau}_{i,2t-1} := \overline{p}_{i,2t-1} \cup \breve{c}_{2t-1}$ for any $t \in [H]$. The new Dec-POMDP $\mathcal{D}'_{\mathcal{L}}$ is not equivalent to $\mathcal{D}^{\dagger}_{\mathcal{L}}$ in general, since it enlarges the strategy space at the odd timesteps. However, if we define new strategy spaces in $\mathcal{D}'_{\mathcal{L}}$ as $\overline{\mathcal{G}}_{i,2t-1} : \overline{\mathcal{C}}_{2t-1} \to \overline{\mathcal{A}}_{i,2t-1}, \overline{\mathcal{G}}_{i,2t} : \overline{\mathcal{T}}_{i,2t} \to \overline{\mathcal{A}}_{i,2t}$ for each $t \in [H], i \in [n]$, and thus define $\overline{\mathcal{G}}_h$ to be the associated joint strategy space, then solving $\mathcal{D}^{\dagger}_{\mathcal{L}}$ is equivalent to finding a *best-in-class* team-optimal strategy of $\mathcal{D}'_{\mathcal{L}}$ within space $\overline{\mathcal{G}}_{1:\overline{H}}$, as shown below.

Theorem IV.6. Let $\mathcal{D}^{\dagger}_{\mathcal{L}}$ be an sQC Dec-POMDP generated from $\mathcal{L}$ after reformulation and strict expansion, and $\mathcal{D}'_{\mathcal{L}}$ be the refinement of $\mathcal{D}^{\dagger}_{\mathcal{L}}$ as introduced above. Then, finding the optimal strategy in $\mathcal{D}^{\dagger}_{\mathcal{L}}$ is equivalent to finding the optimal strategy of $\mathcal{D}'_{\mathcal{L}}$ in the space $\overline{\mathcal{G}}_{1:\overline{H}}$, and $\mathcal{D}'_{\mathcal{L}}$ satisfies the

following information evolution rules: for each $h \in [\overline{H}]$:

$$\overline{c}_h = \overline{c}_{h-1} \cup \overline{z}_h, \quad \overline{z}_h = \overline{\chi}_h(\overline{p}_{h-1}, \overline{a}_{h-1}, \overline{o}_h)$$
$$\text{for each } i \in [n], \quad \overline{p}_{i,h} = \overline{\xi}_{i,h}(\overline{p}_{i,h-1}, \overline{a}_{i,h-1}, \overline{o}_{i,h}),$$

with some functions $\{\overline{\chi}_h\}_{h\in[\overline{H}]}, \{\overline{\xi}_{i,h}\}_{i\in[n],h\in[\overline{H}]}$. Furthermore, if Assumptions III.5 and III.7 hold, then $\mathcal{D}'_{\mathcal{L}}$ has SI-CIBs with respect to the strategy space $\overline{\mathcal{G}}_{1:\overline{H}}$, i.e., for any $h \in [\overline{H}], \overline{s}_h \in \overline{\mathcal{S}}, \overline{p}_h \in \overline{\mathcal{P}}_h, \overline{c}_h \in \overline{\mathcal{C}}_h, \overline{g}_{1:h-1}, \overline{g}'_{1:h-1} \in \overline{\mathcal{G}}_{1:h-1}$ such that $\overline{c}_h$ is reachable under both $\overline{g}_{1:h-1}$ and $\overline{g}'_{1:h-1}$, it holds that

$$\mathbb{P}_h^{\mathcal{D}'_{\mathcal{L}}}(\overline{s}_h, \overline{p}_h \,|\, \overline{c}_h, \overline{g}_{1:h-1}) = \mathbb{P}_h^{\mathcal{D}'_{\mathcal{L}}}(\overline{s}_h, \overline{p}_h \,|\, \overline{c}_h, \overline{g}'_{1:h-1}). \quad \text{(IV.5)}$$

D. Planning in QC LTC with Finite-Time Complexity

Now we focus on how to solve the SI-CIB Dec-POMDP $\mathcal{D}'_{\mathcal{L}}$ without computationally intractable oracles, which has been studied in [14]. Given a Dec-POMDP $\mathcal{D}'_{\mathcal{L}}$ with SI-CIBs, [14] proposed to construct an (ϵ_r, ϵ_z)-*expected*-approximate common information model $\mathcal{M}$ through *finite memory* (as defined in §C), when $\mathcal{D}'_{\mathcal{L}}$ is γ-observable. ϵ_r and ϵ_z here denote the approximation errors for rewards and transitions, respectively, for which we defer a detailed introduction to §C). However, the Dec-POMDP $\mathcal{D}'_{\mathcal{L}}$ obtained from LTC has two key differences from the general ones considered in [14]. First, $\mathcal{D}'_{\mathcal{L}}$ does not satisfy the γ-observability assumption *throughout* the whole $2H$ timesteps. Fortunately, since the emissions at odd steps are still γ-observable, while those at even steps are unimportant as the states remain *unchanged* from the previous step, similar results of *belief contraction* and near-optimality of finite-memory truncation as in [14] can still be obtained. Second, the rewards at the odd steps can now depend on the *private information* $\overline{p}_h$, instead of the state $\overline{s}_h$. Thanks to the approximate common-information-based beliefs defined as $\{\mathbb{P}_h^{\mathcal{M}}(\overline{s}_h, \overline{p}_h \,|\, \widehat{c}_h)\}_{h\in[H]}$, where $\widehat{c}_h$ is the approximate common information compressed from $\overline{c}_h$, which provide the *joint* probability of $\overline{s}_h$ and $\overline{p}_h$, we can still properly evaluate the rewards at the odd steps in the algorithms of [14].

Hence, we can leverage the approaches in [14] to find the optimal strategy $\overline{g}^*_{1:\overline{H}}$ by finding an optimal prescription $\gamma^*_{1:\overline{H}}$ under each possible $\widehat{c}_{1:\overline{H}}$ with backward induction over the timesteps $h = \overline{H}, \cdots, 1$.

Note that in each step of the backward induction, a *Team Decision problem* [29] needs to be solved for each $\widehat{c}_h$, which is known to be `NP-hard` in general [29]:

$$\left(\widehat{g}^*_{1,h}(\cdot \,|\, \widehat{c}_h, \cdot), \cdots, \widehat{g}^*_{n,h}(\cdot \,|\, \widehat{c}_h, \cdot)\right) \leftarrow \operatorname*{argmax}_{\gamma_h} Q_h^{*,\mathcal{M}}(\widehat{c}_h, \gamma_h), \quad \text{(IV.6)}$$

where the Q-value function and the prescriptions γ_h are defined in §C. Hence, to obtain overall computational tractability, we make the following *one-step* tractability assumption, as in [14].

Assumption IV.7 (One-step tractability of $\mathcal{M}$)**.** The one-step Team Decision problems induced by $\mathcal{M}$ (i.e., Eq. (IV.6)) can be solved in polynomial time for all $h = 2t, t \in [H]$.

Several remarks are in order regarding the assumption. First, it can be viewed as a *minimal* assumption when it comes to computational tractability – even with $H = 1$ and no LTC, one-step TDP requires additional structures to be solved efficiently. Second, since the Dec-POMDP here is reformulated from an LTC under Assumption III.4, it suffices to only assume one-step tractability for the *control* (i.e., even) steps. Third, even without Assumption IV.7, the SI-CIB property of $\mathcal{D}'_{\mathcal{L}}$ and thus the derivation of *fixed, tractable size* dynamic programs to solve $\mathcal{L}$ efficiently still hold. Without such efforts, intractably many TDPs may need to be solved, leaving it less hopeful for computational tractability (even under Assumption IV.7). Finally, such an assumption is satisfied for several classes of Dec-POMDPs with information sharing, see §G for more examples. With this assumption, we can obtain a planning algorithm with quasi-polynomial time complexity as follows.

Theorem IV.8. Given any QC LTC problem $\mathcal{L}$ satisfying Assumptions III.1, III.4, III.5, III.7, and IV.7, we can construct an SI Dec-POMDP problem $\mathcal{D}'_{\mathcal{L}}$ such that for any $\epsilon > 0$, solving an ϵ-team optimal strategy in $\mathcal{D}'_{\mathcal{L}}$ can give us an ϵ-team optimal strategy of $\mathcal{L}$, and the following holds. Fix $\epsilon_r, \epsilon_z > 0$ and given any (ϵ_r, ϵ_z)-expected-approximate common information model $\mathcal{M}$ for $\mathcal{D}'_{\mathcal{L}}$ that is consistent with some given approximate belief $\{\mathbb{P}_h^{\mathcal{M},c}(\overline{s}_h, \overline{p}_h \,|\, \widehat{c}_h)\}_{h\in[\overline{H}]}$, Algorithm 1 can compute a $(2\overline{H}\epsilon_r + \overline{H}^2\epsilon_z)$-team optimal strategy for the original LTC problem $\mathcal{L}$ with time complexity $\max_{h\in[\overline{H}]} |\widehat{\mathcal{C}}_h| \cdot \mathtt{poly}(|\mathcal{S}|, |\mathcal{A}_h|, |\mathcal{P}_h|, \overline{H})$. In particular, for fixed $\epsilon > 0$, if $\mathcal{L}$ has any one of baseline sharing protocols as in Appendix A, one can construct a $\mathcal{M}$ and apply Algorithm 1 to compute an ϵ-team optimal strategy for $\mathcal{L}$ in quasi-polynomial time.

E. LTC with Finite-Time and Sample Complexities

Based on the planning results, we are now ready to solve the *learning* problem with both time and sample complexity guarantees. In particular, we can treat the samples from $\mathcal{L}$ as the samples from $\mathcal{D}'_{\mathcal{L}}$: the *reformulation* step (§IV-A) does not change the system dynamics, but only maps the information to different random variables; the *expansion* step (§IV-B) only requires agents to share more actions with each other, without changing the input and output of the environment; the *refinement* step (§IV-C) only recovers the private information the agents had in the original $\mathcal{L}$. This way, we can utilize similar algorithmic ideas in [14] to develop a learning algorithm for LTC problems. See §C for more details of the provable LTC algorithms adapted from [14]. The algorithm has the following finite-time and sample complexity guarantees.

Theorem IV.9. Given any QC LTC problem $\mathcal{L}$ satisfying Assumptions III.1, III.4, III.5, and III.7, we can construct an SI-CIB Dec-POMDP problem $\mathcal{D}'_{\mathcal{L}}$. Moreover, there exists an LTC algorithm (see Algorithm 2 in §C) learning in $\mathcal{D}'_{\mathcal{L}}$, such that if the learned expected-approximate-common-information models $\widehat{\mathcal{M}}$ satisfy Assumption IV.7, then an

ϵ-team-optimal strategy for $\mathcal{L}$ can be learned with high probability, with time and sample complexities polynomial in the parameters of $\widehat{\mathcal{M}}$. Specifically, if $\mathcal{L}$ has the baseline sharing protocols as in §A, then such an algorithm can learn an ϵ-team optimal strategy for $\mathcal{L}$ with high probability, with both quasi-polynomial time and sample complexities.

V. Solving General QC Dec-POMDPs

In §IV, we developed a pipeline for solving a special class of QC Dec-POMDPs generated by LTCs, without computationally intractable oracles. In fact, the pipeline can also be extended to solving general QC Dec-POMDPs, which thus advances the results in [14] that can only address *SI-CIB* Dec-POMDPs, a result of independent interest. Without much confusion given the context, we will adapt the notation for LTCs to studying general Dec-POMDPs: we set $h^+ = h^- = h$ and void the additional sharing protocol. We extend the results in §IV to general QC Dec-POMDPs as follows.

Theorem V.1. Consider a Dec-POMDP $\mathcal{D}$ under Assumptions II.1 (e). If $\mathcal{D}$ is sQC and satisfies Assumptions II.2, III.5, and III.7, then it has SI-CIBs. Meanwhile, if $\mathcal{D}$ has SI-CIBs and perfect recall, then it is sQC (up to null sets).

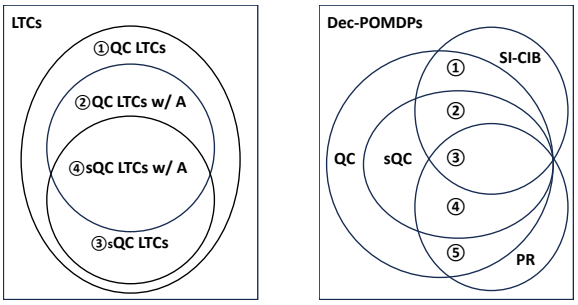
Perfect recall here [22] means that the agents will never forget their own past information and actions (as formally defined in §D). Note that Assumption II.1 (e) is similar but different from perfect recall: it is implied by the latter with $o_{i,h} \in \tau_{i,h}$. Also, Assumptions III.5, III.7, and II.2 were originally made for LTCs, and here we meant to impose them for Dec-POMDPs with $h^+ = h^- = h$. Finally, by sQC up to null sets, we meant that if agent (i_1, h_1) influences agent (i_2, h_2) in the intrinsic model of the Dec-POMDP, then under any strategy $\overline{g}_{1:\overline{H}}$, $\sigma(\overline{\tau}_{i_1,h_1}) \subseteq \sigma(\overline{\tau}_{i_2,h_2})$ except the null sets generated by $\overline{g}_{1:\overline{H}}$, where we add $^-$ for all the notation in the Dec-POMDP. Given Theorem V.1 and the results in §IV, we illustrate the relationship between LTCs and Dec-POMDPs with different assumptions and ISs in Fig. 2, which may be of independent interest.

VI. Experimental Results

For the experiments, we validate both the implementability and performance of our LTC algorithms, and conduct ablation studies for LTCs with different communication costs and horizons. We conduct the experiments in Dectiger and Grid3x3, and the setup details are deferred to §I. The attained average-values are presented in Fig. 3, and the learning curves are shown in Fig. 4. The results show that communication is beneficial for agents to obtain higher values with better sample efficiency. Also, cheaper communication costs can encourage agents to share more information, and jointly achieve a better strategy.

VII. Concluding Remarks

We formalized the learning-to-communicate problem under the Dec-POMDP framework, and proposed a few structural assumptions for LTCs with quasi-classical information structures, violating which can cause computational hardness in general. We then developed provable planning and learning algorithms for QC LTCs. Along the way, we also established some relationship between the strictly quasi-classical information structure and the condition of having strategy-independent common-information-based beliefs, as well as solving general Dec-POMDPs without computationally intractable oracles beyond those with the SI-CIB condition. Our work has opened up many future directions, including the formulation, together with the development of provable planning/learning algorithms, of LTC in non-cooperative (game-theoretic) settings, and the relaxation of (some of) the structural assumptions when it comes to equilibrium computation.



(a) (b)

Fig. 2: (a) Venn diagram of LTCs with different ISs: ① QC LTCs. ② QC LTCs satisfying Assumptions III.4, III.5, and III.7. ③ sQC LTCs. ④ sQC LTCs satisfying Assumptions III.4, III.5, and III.7, whose reformulated Dec-POMDPs have SI-CIB; (b) Venn diagram of general Dec-POMDPs with different ISs. PR denotes perfect recall. We construct examples for each areas in §H.

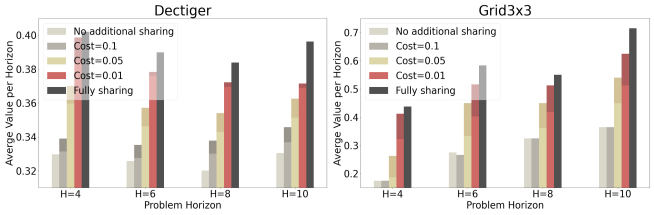


Fig. 3: The average-values achieved under different communication costs and horizons. Each full bar, the dark part, and the light part denote the values associated with the reward, the communication cost, and the overall objective (reward minus cost) of the agents, respectively. Note that, as baselines, there is no communication cost in the *no additional sharing* and *fully sharing* cases.

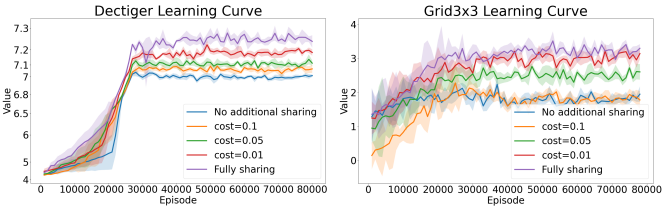


Fig. 4: Learning curves with different communication costs.

REFERENCES

[1] J. Foerster, I. A. Assael, N. De Freitas, and S. Whiteson, "Learning to communicate with deep multi-agent reinforcement learning," in *NeurIPS*, 2016.

[2] S. Sukhbaatar, R. Fergus, *et al.*, "Learning multiagent communication with backpropagation," in *NeurIPS*, 2016.

[3] J. Jiang and Z. Lu, "Learning attentional communication for multi-agent cooperation," in *NeurIPS*, 2018.

[4] S. Tatikonda and S. Mitter, "Control under communication constraints," *IEEE Trans. Autom. Control*, vol. 49, pp. 1056–1068, 2004.

[5] G. N. Nair, F. Fagnani, S. Zampieri, and R. J. Evans, "Feedback control under data rate constraints: An overview," *Proceed. of the IEEE*, vol. 95, pp. 108–137, 2007.

[6] L. Xiao, M. Johansson, H. Hindi, S. Boyd, and A. Goldsmith, "Joint optimization of wireless communication and networked control systems," *Switching and Learning Feedback Sys.*, pp. 248–272, 2005.

[7] S. Yüksel, "Jointly optimal LQG quantization and control policies for multi-dimensional systems," *IEEE Trans. Autom. Control*, vol. 59, pp. 1612–1617, 2013.

[8] S. Sudhakara, D. Kartik, R. Jain, and A. Nayyar, "Optimal communication and control strategies in a multi-agent mdp problem," *arXiv preprint arXiv:2104.10923*, 2021.

[9] D. Kartik, S. Sudhakara, R. Jain, and A. Nayyar, "Optimal communication and control strategies for a multi-agent system in the presence of an adversary," in *IEEE Conf. on Dec. and Control*, 2022.

[10] H. S. Witsenhausen, "Separation of estimation and control for discrete time systems," *Proceed. of the IEEE*, vol. 59, pp. 1557–1566, 1971.

[11] A. Mahajan, N. C. Martins, M. C. Rotkowitz, and S. Yüksel, "Information structures in optimal decentralized control," in *IEEE Conf. on Dec. and Control*, 2012.

[12] S. Yüksel and T. Başar, *Stochastic Teams, Games, and Control under Information Constraints*. Springer Nature, 2023.

[13] D. S. Bernstein, R. Givan, N. Immerman, and S. Zilberstein, "The complexity of decentralized control of markov decision processes," *Math. Oper. Res.*, vol. 27, pp. 819–840, 2002.

[14] X. Liu and K. Zhang, "Partially observable multi-agent reinforcement learning with information sharing," *arXiv preprint arXiv:2308.08705 (short version accepted at ICML 2023)*, 2023.

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

[16] A. Nayyar, A. Gupta, C. Langbort, and T. Başar, "Common information based Markov perfect equilibria for stochastic games with asymmetric information: Finite games," *IEEE Trans. Autom. Control*, vol. 59, pp. 555–570, 2013.

[17] C. H. Papadimitriou and J. N. Tsitsiklis, "The complexity of Markov decision processes," *Math. Oper. Res.*, vol. 12, pp. 441–450, 1987.

[18] C. Lusena, J. Goldsmith, and M. Mundhenk, "Nonapproximability results for partially observable Markov decision processes," *J. Artif. Intell. Res.*, pp. 83–103, 2001.

[19] C. Jin, S. Kakade, A. Krishnamurthy, and Q. Liu, "Sample-efficient reinforcement learning of undercomplete pomdps," in *NeurIPS*, 2020.

[20] Q. Liu, C. Szepesvári, and C. Jin, "Sample-efficient reinforcement learning of partially observable Markov games," in *NeurIPS*, 2022.

[21] A. Altabaa and Z. Yang, "On the role of information structure in reinforcement learning for partially-observable sequential teams and games," in *NeurIPS*, 2024.

[22] H. W. Kuhn, "Extensive games and the problem of information," in *Contrib. Theory Games, Vol. II*. Princeton Univ. Press, 1953.

[23] H. S. Witsenhausen, "The intrinsic model for discrete stochastic control: Some open problems," in *Control Theory, Numer. Methods Comput. Syst. Model., Int. Symp., Rocquencourt*, 1975, pp. 322–335.

[24] A. Mahajan and S. Yüksel, "Measure and cost dependent properties of information structures," in *Amer. Control Conf.*, 2010, pp. 6397–6402.

[25] N. Golowich, A. Moitra, and D. Rohatgi, "Planning and learning in partially observable systems via filter stability," in *Proc. 55th Annu. ACM Symp. Theory Comput.*, 2023.

[26] E. Even-Dar, S. M. Kakade, and Y. Mansour, "The value of observation for monitoring dynamic systems," in *IJCAI*, 2007.

[27] Y.-C. Ho *et al.*, "Team decision theory and information structures in optimal control problems – part i," *IEEE Trans. Autom. Control*, vol. 17, pp. 15–22, 1972.

[28] A. Lamperski and L. Lessard, "Optimal decentralized state-feedback control with sparsity and delays," *Automatica*, pp. 143–151, 2015.

[29] J. Tsitsiklis and M. Athans, "On the complexity of decentralized decision making and detection problems," *IEEE Trans. Autom. Control*, vol. 30, pp. 440–446, 1985.

[30] Q. Liu, A. Chung, C. Szepesvári, and C. Jin, "When is partially observable reinforcement learning not scary?" in *Conference on Learning Theory*. PMLR, 2022, pp. 5175–5220.

[31] N. Golowich, A. Moitra, and D. Rohatgi, "Learning in observable pomdps, without computationally intractable oracles," vol. 35, 2022, pp. 1458–1473.

[32] J. Filar and K. Vrieze, *Competitive Markov decision processes*. Springer, 2012.

[33] Y. Bai and C. Jin, "Provable self-play algorithms for competitive reinforcement learning," in *ICML*, 2020.

[34] J. Peralez, A. Delage, O. Buffet, and J. S. Dibangoye, "Solving hierarchical information-sharing Dec-POMDPs: an extensive-form game approach," *arXiv preprint arXiv:2402.02954*, 2024.

[35] C. Boutilier, "Multiagent systems: Challenges and opportunities for decision-theoretic planning," *AI magazine*, vol. 20, pp. 35–35, 1999.

[36] R. Nair, M. Tambe, M. Yokoo, D. Pynadath, and S. Marsella, "Taming decentralized POMDPs: Towards efficient policy computation for multiagent settings," in *IJCAI*, 2003.

[37] C. Amato, J. Dibangoye, and S. Zilberstein, "Incremental policy generation for finite-horizon Dec-POMDPs," in *Proc. Int. Conf. Autom. Plan. Sched. (ICAPS)*, vol. 19, 2009, pp. 2–9.

APPENDIX

A. Examples of QC LTC

In this section, we introduce 8 examples of QC LTC problems, and 4 of them are extended from the information structures of the baseline sharing protocol considered in the literature [16], [14]. It can be shown that LTC with any of these 8 examples as baseline sharing is QC.

- **Example 1: One-step delayed information sharing:** At timestep $h \in [H]$, agents will share all the action-observation history in the private information until timestep $h-1$. Namely, for any $h \in [H], i \in [n], c_{h^-} = c_{(h-1)^+} \cup \{o_{h-1}, a_{h-1}\}$ and $p_{i,h^-} = \{o_{i,h}\}$.
- **Example 2: State controlled by one controller with asymmetric delayed information sharing:** The state dynamics and reward are controlled by only one agent (without loss of generality, agent 1), i.e., $\mathbb{T}_h(\cdot \,|\, s_h, a_{1,h}, a_{-1,h}) = \mathbb{T}_h(\cdot \,|\, s_h, a_{1,h}, a'_{-1,h})$, $\mathcal{R}_h(\cdot \,|\, s_h, a_{1,h}, a_{-1,h}) = \mathcal{R}_h(\cdot \,|\, s_h, a_{1,h}, a'_{-1,h})$ for all $s_h \in \mathcal{S}, a_{1,h} \in \mathcal{A}_{1,h}, a_{-1,h}, a'_{-1,h} \in \mathcal{A}_{-1,h}$. Agent 1 will share all of her information immediately, while others will share their information with a delay of $d \geq 1$ timesteps[1] in the baseline sharing. Namely, for any $h \in [H], i \neq 1$, $c_{h^-} = c_{(h-1)^+} \cup \{a_{1,h-1}, o_{1,h}, o_{-1,h-d}\}, p_{1,h^-} = \emptyset, p_{i,h^-} = p_{i,(h-1)^+} \cup \{o_{i,h}\} \backslash \{o_{i,h-d}\}$.
- **Example 3: Information sharing with one-directional-one-step-delay:** For convenience, we assume there are 2 agents, and this example can be readily generalized to the multi-agent case. In this case, agent 1 will share the information immediately, while agent 2 will share information with one-step delay. Namely, $c_{1^-} = \{o_{1,1}\}, p_{1,1^-} = \emptyset, p_{2,1^-} = \{o_{2,1}\}$; for any $h \geq 2, i \in [n], c_{h^-} = c_{(h-1)^+} \cup \{o_{1,h}, o_{2,h-1}, a_h\}, p_{1,h^-} = \emptyset, p_{2,h^-} = \{o_{2,h}\}$.
- **Example 4: Uncontrolled state process:** The state transition and reward do not depend on the action of agents, i.e., $\mathbb{T}_h(\cdot \,|\, s_h, a_h) = \mathbb{T}_h(\cdot \,|\, s_h, a'_h), \mathcal{R}_h(\cdot \,|\, s_h, a_h) = \mathcal{R}_h(\cdot \,|\, s_h, a'_h)$ for any $s_h \in \mathcal{S}, a_h, a'_h \in \mathcal{A}_h$. All agents will share their information with a delay of $d \geq 1$. For any $h \in [H], i \in [n]$, $c_{h^-} = c_{(h-1)^+} \cup \{o_{h-d}\}, p_{i,h^-} = p_{i,(h-1)^+} \cup \{o_{i,h}\} \backslash \{o_{i,h-d}\}$.
- **Example 5: One-step delayed observation sharing:** At timestep $h \in [H]$, each agent has access to observations of all agents until timestep $h-1$ and her present observation. Namely, for any $h \in [H], i \in [n], c_{h^-} = c_{(h-1)^+} \cup \{o_{h-1}\}$ and $p_{i,h^-} = \{o_{i,h}\}$.
- **Example 6: One-step delayed observation and two-step delayed control sharing:** At timestep $h \in [H]$, each agent will share the observation history until timestep $h-1$ and action history until timestep $h-2$ from the private information. Namely, for any $h \in [H], i \in [n], c_{h^-} = c_{(h-1)^+} \cup \{o_{h-1}, a_{h-2}\}, p_{i,h^-} = \{o_{i,h}, a_{i,h-1}\}$.
- **Example 7: State controlled by one controller with asymmetric delayed observation sharing:** The state dynamics and reward are controlled by only one agent (i.e., system dynamics are the same as **Example 2**). Agent 1 will share all of her observations immediately, while others will share their observations with a delay of $d \geq 1$ timesteps in baseline sharing. Namely, for any $h \in [H], i \neq 1, c_{h^-} = c_{(h-1)^+} \cup \{o_{1,h}, o_{-1,h-d}\}, p_{1,h^-} = \emptyset, p_{i,h^-} = p_{i,(h-1)^+} \cup \{o_{i,h}\} \backslash \{o_{i,h-d}\}$.
- **Example 8: State controlled by one controller with asymmetric delayed observation and two-step delayed action sharing:** The state dynamics and reward are controlled by only one agent (i.e., system dynamics are the same as **Example 2**). At timestep $h \in [H]$, agent 1 will share all of her observations immediately and her action history until timestep $h-2$, while others will share their observations with a delay of $d \geq 1$. Namely, for any $h \in [H], i \neq 1, c_{h^-} = c_{(h-1)^+} \cup \{o_{1,h}, a_{1,h-2}, o_{-1,h-d}\}, p_{1,h^-} = \{a_{1,h-1}\}, p_{i,h^-} = p_{i,(h-1)^+} \cup \{o_{i,h}\} \backslash \{o_{i,h-d}\}$.

For **Examples 1, 5, 6**, for computational considerations later (see §IV), we may additionally assume that for any $h \in [H]$, the state s_h can be partitioned into n local states as $s_h = (s_{1,h}, s_{2,h}, \cdots, s_{n,h})$, and the transition kernel and observation emission have the **factorized** forms of $\mathbb{T}_h(s_{h+1} \,|\, s_h, a_h) = \prod_{i=1}^n \mathbb{T}_{i,h}(s_{i,h+1} \,|\, s_{i,h}, a_{i,h}), \mathbb{O}_h(o_h \,|\, s_h) = \prod_{i=1}^n \mathbb{O}_{i,h}(o_{i,h} \,|\, s_{i,h})$. Furthermore, the communication cost and reward functions may be assumed to be decoupled as $\mathcal{K}_h(p_h, m_h) = \sum_{i=1}^n \mathcal{K}_{i,h}(p_{i,h}, m_{i,h}), \mathcal{R}_h(s_h, a_h) = \sum_{i=1}^n \mathcal{R}_{i,h}(s_{i,h}, a_{i,h})$.

Remark .1. These additional conditions on $\mathbb{T}_h, \mathbb{O}_h, \mathcal{K}_h, \mathcal{R}_h$ are used to ensure the one-step tractability of the backward induction for solving the LTC problem (see Assumption IV.7). Note that these **Examples 1, 5, 6** are QC even without these additional conditions. Moreover, note that **Examples 2, 3, 4, 7, 8** above automatically satisfy Assumption IV.7 (see the proofs in §C for more details).

In fact, the first 4 examples are all sQC LTC problems, while the other 4 examples are QC but not sQC problems, as shown in the following lemma.

Lemma .2. Given an LTC problem $\mathcal{L}$. If the baseline sharing of $\mathcal{L}$ is one of the first 4 examples above, then $\mathcal{L}$ is sQC. If the baseline sharing of $\mathcal{L}$ is one of the last 4 examples above, then $\mathcal{L}$ is QC but not sQC.

[1]Throughout this paper, we view the delay d as a *constant*, although our final bounds in §IV-D and §IV-E also apply for $d = \text{poly} \log H$. See the proofs in §C for more discussions.

Proof. Let $\overline{\mathcal{D}}_{\mathcal{L}}$ denote the Dec-POMDP induced by $\mathcal{L}$ (see Definition II.4). We prove this lemma case by case. For convenience, we use $\dot{\cdot}$ for the notation of the elements in $\overline{\mathcal{D}}_{\mathcal{L}}$.

- **Example 1:** The information in $\overline{\mathcal{D}}_{\mathcal{L}}$ evolves as $\forall h \in [H], i \in [n], \dot{c}_h = \{\dot{o}_{1:h-1}, \dot{a}_{1:h-1}\}$ and $\dot{p}_{i,h} = \{\dot{o}_{i,h}\}$. Then, for any $i_1, i_2 \in [n], h_1, h_2 \in [H], h_1 < h_2$, $\dot{\tau}_{i_1,h_1} = \{\dot{o}_{1:h_1-1}, \dot{a}_{1:h_1-1}, \dot{o}_{i_1,h_1}\} \subseteq \dot{c}_{h_1+1} \subseteq \dot{c}_{h_2} \subseteq \dot{\tau}_{i_2,h_2}$, and $\dot{a}_{i_1,h_1} \subseteq \dot{c}_{h_1+1} \subseteq \dot{c}_{h_2} \subseteq \dot{\tau}_{i_2,h_2}$. Therefore, we have $\sigma(\dot{\tau}_{i_1,h_1}) \subseteq \sigma(\dot{\tau}_{i_2,h_2})$, and thus $\mathcal{L}$ is sQC.
- **Example 2:** The information in $\overline{\mathcal{D}}_{\mathcal{L}}$ evolves as $\forall h \in [H], i \neq 1, \dot{c}_h = \{\dot{a}_{1,1:h-1}, \dot{o}_{1,1:h-1}, \dot{o}_{-1,1:h-d}\}, \dot{p}_{1,h} = \emptyset, \dot{p}_{i,h} = \{o_{i,h-d+1:h}\}$. Then, for any $i_1, i_2 \in [n], h_1, h_2 \in [H], h_1 < h_2$. If $i_1 \neq 1$, then agent (i_1, h_1) will not influence agent (i_2, h_2). If $i_1 = 1$, then $\dot{\tau}_{i_1,h_1} = \{\dot{o}_{1,1:h_1}, \dot{a}_{1,1:h_1-1}, \dot{o}_{-1,1:h_1-d}\} \subseteq \dot{c}_{h_1+1} \subseteq \dot{c}_{h_2} \subseteq \dot{\tau}_{i_2,h_2}$, and $\dot{a}_{i_1,h_1} \subseteq \dot{c}_{h_1+1} \subseteq \dot{c}_{h_2} \subseteq \dot{\tau}_{i_2,h_2}$. Therefore, we have $\sigma(\dot{\tau}_{i_1,h_1}) \subseteq \sigma(\dot{\tau}_{i_2,h_2})$ if agent (i_1, h_1) influences agent (i_2, h_2), and thus $\mathcal{L}$ is sQC.
- **Example 3:** The information in $\overline{\mathcal{D}}_{\mathcal{L}}$ evolves as $\forall h \in [H], \dot{c}_h = \{\dot{o}_{1:h-1}, \dot{a}_{1:h-1}, \dot{o}_{1,h}\}$ and $\dot{p}_{1,h} = \emptyset, \dot{p}_{2,h} = \{\dot{o}_{i,h}\}$. Then, for any $i_1, i_2 \in [n], h_1, h_2 \in [H], h_1 < h_2$, $\dot{a}_{i_1,h_1} \subseteq \dot{c}_{h_1+1} \subseteq \dot{c}_{h_2} \subseteq \dot{\tau}_{i_2,h_2}$. If $i_1 = 1$, then $\dot{\tau}_{i_1,h_1} = \{\dot{o}_{1:h_1-1}, \dot{a}_{1:h_1-1}, \dot{o}_{1,h_1}\} \subseteq \dot{c}_{h_1+1} \subseteq \dot{c}_{h_2} \subseteq \dot{\tau}_{i_2,h_2}$. If $i_1 = 2$, then $\dot{\tau}_{i_1,h_1} = \{\dot{o}_{1:h_1}, \dot{a}_{1:h_1-1}\} \subseteq \dot{c}_{h_1+1} \subseteq \dot{c}_{h_2} \subseteq \dot{\tau}_{i_2,h_2}$. Therefore, we have $\sigma(\dot{\tau}_{i_1,h_1}) \subseteq \sigma(\dot{\tau}_{i_2,h_2})$, and thus $\mathcal{L}$ is sQC.
- **Example 4:** Since in $\overline{\mathcal{D}}_{\mathcal{L}}$, for any $i_1, i_2 \in [n], h_1, h_2 \in [H]$, agent (i_1, h_1) does not influence agent (i_2, h_2), then $\mathcal{L}$ is sQC.
- **Example 5:** The information in $\overline{\mathcal{D}}_{\mathcal{L}}$ evolves as $\forall h \in [H], i \in [n], \dot{c}_h = \{\dot{o}_{1:h-1}\}$ and $\dot{p}_{i,h} = \{\dot{o}_{i,h}\}$. Then, for any $i_1, i_2 \in [n], h_1, h_2 \in [H], h_1 < h_2$, $\dot{\tau}_{i_1,h_1} = \{\dot{o}_{1:h_1-1}, \dot{o}_{i_1,h_1}\} \subseteq \dot{c}_{h_1+1} \subseteq \dot{c}_{h_2} \subseteq \dot{\tau}_{i_2,h_2}$. However, agent $(1,1)$ may influence agent $(1,2)$ but $\sigma(\dot{a}_{1,1}) \nsubseteq \sigma(\dot{\tau}_{1,2})$. Hence, $\mathcal{L}$ is QC but not sQC.
- **Example 6:** The information in $\overline{\mathcal{D}}_{\mathcal{L}}$ evolves as $\forall h \in [H], i \in [n], \dot{c}_h = \{\dot{o}_{1:h-1}, \dot{a}_{1:h-2}\}$ and $\dot{p}_{i,h} = \{\dot{o}_{i,h}, \dot{a}_{i,h-1}\}$. Then, for any $i_1, i_2 \in [n], h_1, h_2 \in [H], h_1 < h_2$, $\dot{\tau}_{i_1,h_1} = \{\dot{o}_{1:h_1-1}, \dot{a}_{1:h_1-2}, \dot{o}_{i_1,h_1}, \dot{a}_{i_1,h_1-1}\} \subseteq \dot{c}_{h_1+1} \subseteq \dot{c}_{h_2} \subseteq \dot{\tau}_{i_2,h_2}$, and $\dot{a}_{i_1,h_1} \subseteq \dot{c}_{h_1+1} \subseteq \dot{c}_{h_2} \subseteq \dot{\tau}_{i_2,h_2}$. However, agent $(1,1)$ may influence agent $(2,2)$ but $\sigma(\dot{a}_{1,1}) \nsubseteq \sigma(\dot{\tau}_{2,2})$. Hence, $\mathcal{L}$ is QC but not sQC.
- **Example 7:** The information in $\overline{\mathcal{D}}_{\mathcal{L}}$ evolves as $\forall h \in [H], i \neq 1, \dot{c}_h = \{\dot{o}_{1,1:h-1}, \dot{o}_{-1,1:h-d}\}, \dot{p}_{1,h} = \emptyset, \dot{p}_{i,h} = \{o_{i,h-d+1:h}\}$. Then, for any $i_1, i_2 \in [n], h_1, h_2 \in [H], h_1 < h_2$. If $i_1 \neq 1$, then agent (i_1, h_1) will not influence agent (i_2, h_2). If $i_1 = 1$, then $\dot{\tau}_{i_1,h_1} = \{\dot{o}_{1,1:h_1}, \dot{o}_{-1,1:h_1-d}\} \subseteq \dot{c}_{h_1+1} \subseteq \dot{c}_{h_2} \subseteq \dot{\tau}_{i_2,h_2}$. Therefore, we have $\sigma(\dot{\tau}_{i_1,h_1}) \subseteq \sigma(\dot{\tau}_{i_2,h_2})$ if agent (i_1, h_1) influences agent (i_2, h_2). However, agent $(1,1)$ may influence agent $(1,2)$ but $\sigma(\dot{a}_{1,1}) \nsubseteq \sigma(\dot{\tau}_{1,2})$. Hence, $\mathcal{L}$ is QC but not sQC.
- **Example 8:** The information in $\overline{\mathcal{D}}_{\mathcal{L}}$ evolves as $\forall h \in [H], i \neq 1, \dot{c}_h = \{\dot{o}_{1,1:h-1}, \dot{a}_{1,1:h-2}, \dot{o}_{-1,1:h-d}\}, \dot{p}_{1,h} = \{\dot{a}_{1,h-1}\}, \dot{p}_{i,h} = \{o_{i,h-d+1:h}\}$. Then, for any $i_1, i_2 \in [n], h_1, h_2 \in [H], h_1 < h_2$. If $i_1 \neq 1$, then agent (i_1, h_1) will not influence agent (i_2, h_2). If $i_1 = 1$, then $\dot{\tau}_{i_1,h_1} = \{\dot{o}_{1,1:h_1}, \dot{a}_{1,h_1-1}, \dot{o}_{-1,1:h_1-d}\} \subseteq \dot{c}_{h_1+1} \subseteq \dot{c}_{h_2} \subseteq \dot{\tau}_{i_2,h_2}$. Therefore, we have $\sigma(\dot{\tau}_{i_1,h_1}) \subseteq \sigma(\dot{\tau}_{i_2,h_2})$ if agent (i_1, h_1) influences agent (i_2, h_2). However, agent $(1,1)$ may influence agent $(2,2)$ but $\sigma(\dot{a}_{1,1}) \nsubseteq \sigma(\dot{\tau}_{2,2})$. Hence, $\mathcal{L}$ is QC but not sQC.

This completes the proof. □

B. Deferred Details of §III

Remark .3. In the following proofs, for clarity, we use $O, A, M, C, P, \mathscr{T}$ to denote the realizations of random variables o, a, m, c, p, τ with the same subscripts.

As a preliminary, we first have the following lemma.

Lemma .4. Given any QC LTC $\mathcal{L}$, its induced Dec-POMDP $\overline{\mathcal{D}}_{\mathcal{L}}$, and any $i_1, i_2 \in [n], h_1, h_2 \in [H]$. If agent (i_1, h_1) influences agent (i_2, h_2) in the intrinsic model of $\overline{\mathcal{D}}_{\mathcal{L}}$, then for the random variables $\tau_{i_1,h_1^-}, \tau_{i_2,h_2^-}$ in $\mathcal{L}$, we have $\sigma(\tau_{i_1,h_1^-}) \subseteq \sigma(\tau_{i_2,h_2^-})$. Moreover, if $\mathcal{L}$ is sQC, then for random variables $a_{i_1,h_1}, \tau_{i_2,h_2^-}$ in $\mathcal{L}$, we have $\sigma(a_{i_1,h_1}) \subseteq \sigma(\tau_{i_2,h_2^-})$.

Proof. We denote by $\check{\tau}_{i_1,h_1}, \check{\tau}_{i_2,h_2}$ the information of agent $(i_1, h_1), (i_2, h_2)$ in the problem $\overline{\mathcal{D}}_{\mathcal{L}}$. From the definition of $\overline{\mathcal{D}}_{\mathcal{L}}$ being QC, if agent (i_1, h_1) influences agent (i_2, h_2), then $\sigma(\check{\tau}_{i_1,h_1}) \subseteq \sigma(\check{\tau}_{i_2,h_2})$. Since for any $h \in [H], i \in [n], \check{\tau}_{i,h}$ is the information of agent (i, h) without additional sharing, then we know that $\tau_{i,h^-} \backslash \check{\tau}_{i,h} \subseteq \cup_{t=1}^{h-1} z_t^a, \tau_{i,h^+} \backslash \check{\tau}_{i,h} \subseteq \cup_{t=1}^{h} z_t^a$. Therefore, we know that $\sigma(\tau_{i_1,h_1^-} \backslash \check{\tau}_{i_1,h_1}) \subseteq \sigma(\cup_{t=1}^{h-1} z_t^a) \subseteq \sigma(c_{h_1^-}) \subseteq \sigma(c_{h_2^-}) \subseteq \sigma(\tau_{i_2,h_2^-})$. Also, we know $\sigma(\check{\tau}_{i_1,h_1}) \subseteq \sigma(\check{\tau}_{i_2,h_2}) \subseteq \sigma(\tau_{i_2,h_2^-})$. Thus, we can conclude that $\sigma(\tau_{i_1,h_1^-}) \subseteq \sigma(\tau_{i_2,h_2^-})$. Moreover, if $\mathcal{L}$ is sQC, then from the the definition of $\overline{\mathcal{D}}_{\mathcal{L}}$ being sQC and agent (i_1, h_1) influences agent (i_2, h_2) in $\overline{\mathcal{D}}_{\mathcal{L}}$, it holds that $\sigma(a_{i_1,h_1}) \subseteq \sigma(\check{\tau}_{i_2,h_2}) \subseteq \sigma(\tau_{i_2,h_2^-})$. □

1) Proof of Lemma III.2:

Proof. We first have the following proposition on the hardness of solving POMDPs.

Proposition .5. There exists an $\epsilon > 0$, such that computing an ϵ-additive optimal strategy in POMDPs is `PSPACE-hard`.

One can adapt the proof of [18, Theorem 4.11], which proved the `PSPACE-hardness` of computing an *ϵ-relative* optimal strategy in POMDPs, to obtain such a result for an *ϵ-additive* one. In particular, one can verify that any ϵ-additive optimal strategy in the POMDP with bounded reward $[0, 1]$ can give an ϵ-relative optimal strategy in the

POMDP constructed in the proof of Theorem 4.11 therein. Therefore, computing ϵ-additive optimal strategy in POMDPSs is PSPACE-hard. Now we proceed with the proof of Lemma III.2 based on Proposition .5. Given any POMDP $\mathcal{P} = (\mathcal{S}^{\mathcal{P}}, \mathcal{A}^{\mathcal{P}}, \mathcal{O}^{\mathcal{P}}, \{\mathbb{O}_h^{\mathcal{P}}\}_{h\in[H^{\mathcal{P}}]}, \{\mathbb{T}_h^{\mathcal{P}}\}_{h\in[H^{\mathcal{P}}]}, \{\mathcal{R}_h^{\mathcal{P}}\}_{h\in[H^{\mathcal{P}}]}, \mu_1^{\mathcal{P}})$, we can construct an LTC $\mathcal{L}$ as follows:

- Number of agents: $n = 3$; length of episode: $H = H^{\mathcal{P}}$.
- Underlying state space: $\mathcal{S} = \mathcal{S}^{\mathcal{P}} \times [2]$. For any $s \in \mathcal{S}$, we can split $s = (s^1, s^2)$, where $s^1 \in \mathcal{S}^{\mathcal{P}}, s^2 \in [2]$. Intial state distribution: $\forall s \in \mathcal{S}, \mu_1(s) = \mu_1^{\mathcal{P}}(s^1)/2$.
- Control action space: For any $h \in [H], \mathcal{A}_{1,h} = \mathcal{A}^{\mathcal{P}}, \mathcal{A}_{2,h} = [2], \mathcal{A}_{3,h} = \{\emptyset\}$.
- Transition functions: For any $h \in [H], s_h, s_{h+1} \in \mathcal{S}, a_h \in \mathcal{A}_h, \mathbb{T}_h(s_{h+1} \,|\, s_h, a_h) = \mathbb{T}_h^{\mathcal{P}}(s_{h+1}^1 \,|\, s_h^1, a_{1,h})\mathbb{1}[s_{h+1}^2 = a_{2,h}]$.
- Observation space: For any $h \in [H], \mathcal{O}_{1,h} = \mathcal{O}_t^{\mathcal{P}} \times [2], \mathcal{O}_{2,h} = \mathcal{O}_{3,h} = \mathcal{S}$. For nay $o_{1,h} \in \mathcal{O}_{1,h}$, similarly, we can split $o_{1,h} = (o_{1,h}^1, o_{1,h}^2)$, where $o_{1,h}^1 \in \mathcal{O}_h^{\mathcal{P}}, o_{1,h}^2 \in [2]$.
- Emission matrix: For any $h \in [H], o_h \in \mathcal{O}_h, s_h \in \mathcal{S}, \mathbb{O}_h(o_h \,|\, s_h) = \mathbb{O}_h^{\mathcal{P}}(o_{1,h}^1 \,|\, s_h^1)\mathbb{1}[o_{1,h}^2 = s_h^2, o_{2,h} = o_{3,h} = s_h]$.
- The baseline sharing: null.
- The communication action space: For any $h \in [H], \mathcal{M}_{1,h} = \mathcal{M}_{2,h} = \{0,1\}^{2h-1}, \mathcal{M}_{3,h} = \{0,1\}^h$. For any $i \in [2], p_{i,h^-} \in \mathcal{P}_{i,h^-}, \phi_{i,h}(p_{i,h^-}, m_{i,h}) = \{o_{i,k} \,|\, k \le h, (2k-1)$-th digit of p_{i,h^-} is 1 and $o_{i,k} \in p_{i,h^-}\} \cup \{a_{i,k} \,|\, k \le h, 2k$-th digit of p_{i,h^-} is 1 and $a_{i,k} \in p_{i,h^-}\} \cup \{m_{i,h}\}$. For agent $3, p_{3,h^-} \in \mathcal{P}_{3,h^-}, \phi_{3,h}(p_{3,h^-}, m_{3,h}) = \{o_{3,k} \,|\, k \le h, k$-th digit of p_{3,h^-} is 1 and $o_{3,k} \in p_{3,h^-}\} \cup \{m_{3,h}\}$.
- Reward function: For any $h \in [H], i \in [3], s_h \in \mathcal{S}, a_h \in \mathcal{A}_h, \mathcal{R}_h(s_h, a_h) = \mathcal{R}_h^{\mathcal{P}}(s_h^1, a_h^1) + \mathbb{1}[a_{2,h} = 1]$.
- Communication cost function: For any $h \in [H], z_h^a \in \mathcal{Z}_h^a, \mathcal{K}_h(z_h^a) = \mathbb{1}[z_h^a \neq \{m_h\}]$. It means that the communication cost is 1 unless there is no additional sharing.
- We restrict the communication strategy only to use c_h as input. And for any $t \in [H-1]$, we remove $a_{3,t}$ in τ_h for any $h > t$.

We first verify that such a construction satisfies Assumptions III.1, III.4, III.5, and III.7.

- $\mathcal{L}$ satisfies Assumption III.1, III.7 because both agent 2 and agent 3 have individual γ-observability with $\gamma = 1$. That is, for any $b_1, b_2 \in \Delta(\mathcal{S}), i = 2, 3$, we have

$$\begin{aligned}
||\mathbb{O}_{i,h}^\top(b_1 - b_2)||_1 &= \sum_{o_{i,h}\in\mathcal{O}_h} |\sum_{s_h\in\mathcal{S}} (b_1(s_h) - b_2(s_h))\mathbb{O}_{i,h}(o_{i,h} \,|\, s_h)| \\
&= \sum_{o_{i,h}\in\mathcal{O}_h} |\sum_{s_h\in\mathcal{S}} (b_1(s_h) - b_2(s_h))\mathbb{1}[o_{i,h} = s_h]| \\
&= \sum_{o_{i,h}\in\mathcal{O}_h} |b_1(o_{i,h}) - b_2(o_{i,h})| = ||b_1 - b_2||_1.
\end{aligned}$$

- $\mathcal{L}$ satisfies Assumption III.4 because we restrict that the communication strategy can only use c_h as input.
- $\mathcal{L}$ satisfies Assumption III.5 since only $a_{3,t}, t \in [H-1]$ do not influence the underlying state, and we remove $a_{3,t}$ from τ_h for any $h > t$.

In this LTC problem $\mathcal{L}$, we can decouple the reward as $r_h = r_{1,h} + r_{2,h}$ with $r_{1,h} = \mathcal{R}_h^{\mathcal{P}}(s_h^1, a_h^1), r_{2,h} = \mathbb{1}[a_{2,h} = 1]$. Agent 2 will always choose $a_{i,h} = 1$ to maximize $r_{2,h} = 1$, and agent 1 needs to maximize $\sum_{h=1}^H r_{1,h}$. There will be no additional sharing since any additional sharing at timestep h will incur a communication cost $\kappa_h = 1 > \max \sum_{t=1}^H r_{1,h}$, and thus it cannot achieve optimum. Therefore, state $s_h^2, h \in [H]$ are dummy states, and agents $2, 3$ are dummy agents. Then, any $(g_{1:H}^{a,*}, g_{1:H}^{m,*})$ being an $\frac{\epsilon}{H}$-team optimal strategy of $\mathcal{L}$ will directly give an ϵ-team-optimal strategy of $\mathcal{P}$ as $\{g_{1,h}^{a,*}\}_{h\in[H]}$. From Proposition .5, we can complete the proof. □

2) Proof of Lemma III.3:

Proof. We prove this result by showing a reduction from the Team Decision problem [29].

Definition .6 (Team decision problem (TDP))**.** Given finite sets Y_1, Y_2, U_1, U_2, a rational probability mass function $p : Y_1 \times Y_2 \to \mathbb{Q}$, and an integer cost function $c : Y_1 \times Y_2 \times U_1 \times U_2 \to \mathbb{N}$, find decision rules $\gamma_i : Y_i \to U_i, i = 1, 2$ that minimize the expected cost

$$J(\gamma_1, \gamma_2) = \sum_{y_1\in Y_1, y_2\in Y_2} c(y_1, y_2, \gamma_1(y_1), \gamma_2(y_2))p(y_1, y_2). \tag{.1}$$

We show the NP-hardness of solving LTC from the problem TDP. Given any TDP $\mathcal{TD} = (\widetilde{Y}_1, \widetilde{Y}_2, \widetilde{U}_1, \widetilde{U}_2, \widetilde{c}, \widetilde{p}, \widetilde{J})$ with $|\widetilde{U}_1| = |\widetilde{U}_2| = 2$, let $\widetilde{U}_1 = \{1, 2\}, \widetilde{U}_2 = \{1, 2\}$, then we can construct an $H = 4$ and 2-agent LTC $\mathcal{L}$ with two parameters $n_1 \in \mathbb{N}, \alpha_1 \in \mathbb{R}, \alpha_2 \in (0, 1)$ (to be specified later) such that:

- Number of agents: $n = 2$; length of episode: $H = 4$.
- Underlying state: $\mathcal{S} = [2]^4$. For each $s_1 \in \mathcal{S}$, we can split s_1 into 4 parts as $s_1 = (s_1^1, s_1^2, s_1^3, s_1^4)$, where $s_1^1, s_1^2, s_1^3, s_1^4 \in [2]$. Similarly, $s_2, s_3, s_4 \in \mathcal{S}$ can be split in the same way.

- Initial state distribution: $\forall s_1 \in \mathcal{S}, \mu_1(s_1) = \frac{1}{16}$.
- Control action space: For the first 2 timesteps, $\forall i = 1, 2, \mathcal{A}_{i,1} = \mathcal{A}_{i,2} = \{\emptyset\}$; for $h = 3, \mathcal{A}_{1,3} = [2], \mathcal{A}_{2,3} = \{\emptyset\}$; for $h = 4, \mathcal{A}_{2,4} = [2], \mathcal{A}_{1,4} = \{\emptyset\}$.
- Transition: $\forall s \in \mathcal{S}, a_1 \in \mathcal{A}_1, a_2 \in \mathcal{A}_2, a_3 \in \mathcal{A}_3, a_4 \in \mathcal{A}_4, \mathbb{T}_1(s \mid s, a_1) = \mathbb{T}_2(s \mid s, a_2) = \mathbb{T}_3(s \mid s, a_3) = \mathbb{T}_4(s \mid s, a_4) = 1$. Note that under the transition dynamics above, $s_1 = s_2 = s_3 = s_4$ always holds, for any $s_1 \in \mathcal{S}$.
- Observation space: $\mathcal{O}_{1,1} = \mathcal{O}_{2,1} = \mathcal{O}_{1,2} = \mathcal{O}_{2,2} = [2] \times \mathcal{S}$, $\mathcal{O}_{1,3} = \widetilde{Y_1} \times \mathcal{S}, \mathcal{O}_{2,3} = \widetilde{Y_2} \times \mathcal{S}, \mathcal{O}_{1,4} = \mathcal{O}_{2,4} = \mathcal{S}$; For each $i \in [2], h \in [2], o_{i,h} \in \mathcal{O}_{i,h}$, we can split $o_{i,h}$ into 2 parts as $o_{i,h} = (o^1_{i,h}, o^2_{i,h})$, where $o^1_{i,h} \in [2], o^2_{i,h} \in \mathcal{S}$. For each $i \in [n], o_{i,3} \in \mathcal{O}_{i,3}$, similarly, we can split $o_{i,3}$ into 2 parts as $o_{i,3} = (o^1_{i,3}, o^2_{i,3})$, where $o^1_{i,3} \in \widetilde{Y}_i, o^2_{i,3} \in \mathcal{S}$.
- The baseline sharing is null.
- Communication action space: For $i \in [2], h \in \{1, 2, 4\}, \mathcal{M}_{i,h} = \{0, 1\}^h, \mathcal{M}_{i,3} = \{1, 2\}$; For each $i \in [2], \phi_{i,h}$ is defined as $\forall h \in \{1, 2, 4\}, \phi_{i,h}(p_{i,h^-}, m_{i,h}) = \{o_{i,k} \mid k \le h, k\text{-th digit of } m_{i,h} \text{ is 1 and } o_{i,k} \in p_{i,h^-}\}$; For $h = 3$, if $m_{i,3} = 1$, then $\phi_{i,h}(p_{i,3^-}, m_{i,3}) = \{o_{i,1}, o_{i,3}, m_{i,3}\}$; if $m_{i,3} = 2$, then $\phi_{i,h}(p_{i,h^-}, m_{i,3}) = \{o_{i,2}, o_{i,3}, m_{i,3}\}$.
- Emission matrix: For any $i \in [2], h \in [2], s_h \in \mathcal{S}, o_{i,h} \in \mathcal{O}_{i,h}, \mathbb{O}_h(o_h \mid s_h) = \Pi_{i=1}^2 \mathbb{O}_{i,h}(o_{i,h} \mid s_h)$ and $\mathbb{O}_{i,h}(o_{i,h} \mid s_h)$ is defined as:

$$\mathbb{O}_{i,h}(o_{i,h} \mid s_h) = \begin{cases} \frac{1-\alpha_2}{16} & o^1_{i,h} = s_h^{i+2h-2}, o^2_{i,h} \neq s_h \\ \frac{1-\alpha_2}{16} + \alpha_2 & o^1_{i,h} = s_h^{i+2h-2}, o^2_{i,h} = s_h \\ 0 & \text{o.w.} \end{cases}.$$

For $i \in [2], s_3 \in \mathcal{S}, o_3 \in \mathcal{O}_3, \mathbb{O}_3(o_3 \mid s_3) = \mathbb{O}^1_3(o^1_3 \mid s_3)\mathbb{O}^2_3(o^2_3 \mid s_3), \mathbb{O}^2_3 = \Pi_{i=1}^2 \mathbb{O}^2_{i,3}(o^2_{i,3} \mid s_3)$ is defined as:

$$\mathbb{O}^1_3(o^1_3 \mid s_3) = \widetilde{p}(o^1_{1,3}, o^1_{2,3})$$
$$\mathbb{O}^2_{i,3}(o^2_3 \mid s_3) = \begin{cases} \frac{1-\alpha_2}{16} & o^2_{i,3} \neq s_3 \\ \frac{1-\alpha_2}{16} + \alpha_2 & o^2_{i,3} = s_3 \end{cases}.$$

And for $i \in [2], s_4 \in \mathcal{S}, o_{i,4} \in \mathcal{O}_{i,4}, \mathbb{O}_4(o_4 \mid s_h) = \Pi_{i=1}^2 \mathbb{O}_{i,4}(o_{i,4} \mid s_4)$ and $\mathbb{O}_{i,4}(o_{i,4} \mid s_4)$ is defined as:

$$\mathbb{O}_{i,4}(o_{i,4} \mid s_4) = \begin{cases} \frac{1-\alpha_2}{16} & o_{i,4} \neq s_4 \\ \frac{1-\alpha_2}{16} + \alpha_2 & o_{i,4} = s_4 \end{cases}.$$

Such an emission matrix means that for each $h \in [2]$ and $i \in [2]$, agent i will accurately observe part of the underlying state s_h^{i+2h-2} and vaguely observe the whole underlying state s_h. For $h = 4, i \in [2]$, agent i can only vaguely observe the whole underlying state s_h. Such a design makes the problem satisfy Assumption III.1. The reward functions are defined as:

$$\mathcal{R}_1(s_1, a_1) = \mathcal{R}_2(s_2, a_2) = 0, \quad \forall s_1, s_2 \in \mathcal{S}, a_1 \in \mathcal{A}_1, a_2 \in \mathcal{A}_2;$$
$$\mathcal{R}_3(s_3, a_3) = \begin{cases} 1 & \text{if } a_{1,3} = s_3^2 \text{ or } a_{1,3} = s_3^4 \\ 0 & \text{o.w.} \end{cases};$$
$$\mathcal{R}_4(s_4, a_4) = \begin{cases} 1 & \text{if } a_{2,4} = s_4^1 \text{ or } a_{2,4} = s_4^3 \\ 0 & \text{o.w.} \end{cases}.$$

The communication cost functions are defined as:

$$\forall h \in \{1, 2, 4\}, z^a_h \in \mathcal{Z}^a_h, \mathcal{K}_h(z^a_h) = 1 \text{ if } z^a_h \neq \{m_{1,h}, m_{2,h}\} \text{ else } 0;$$
$$\mathcal{K}_3(z^a_3) = \begin{cases} \widetilde{c}(o^1_{1,3}, o^1_{2,3}, 1, 1)/\alpha_1 & \text{if } \{o_{1,1}, o_{2,1}\} \subseteq z^a_3 \text{ and } \{o_{1,2}, o_{2,2}\} \cap z^a_3 = \emptyset \\ \widetilde{c}(o^1_{1,3}, o^1_{2,3}, 2, 1)/\alpha_1 & \text{if } \{o_{1,2}, o_{2,1}\} \subseteq z^a_3 \text{ and } \{o_{1,1}, o_{2,2}\} \cap z^a_3 = \emptyset \\ \widetilde{c}(o^1_{1,3}, o^1_{2,3}, 1, 2)/\alpha_1 & \text{if } \{o_{1,1}, o_{2,2}\} \subseteq z^a_3 \text{ and } \{o_{1,2}, o_{2,1}\} \cap z^a_3 = \emptyset \\ \widetilde{c}(o^1_{1,3}, o^1_{2,3}, 2, 2)/\alpha_1 & \text{if } \{o_{1,2}, o_{2,2}\} \subseteq z^a_3 \text{ and } \{o_{1,1}, o_{2,1}\} \cap z^a_3 = \emptyset \end{cases}.$$

Let $\alpha_0 = \max_{y_1, y_2, u_1, u_2} \widetilde{c}(y_1, y_2, u_1, u_2)$, and set $\alpha_1 = 2\alpha_0$. Under such a construction, $\mathcal{L}$ satisfies the following conditions:

- Problem $\mathcal{L}$ is QC: For all $i_1, i_2 \in [2], h_1, h_2 \in [4]$, agent (i_1, h_1) does not influence (i_2, h_2) because agent (i_1, h_1) cannot influence the observation of agent (i_2, h_2), and baseline sharing is null.
- Problem $\mathcal{L}$ satisfies Assumptions III.1 and III.7: We prove this by showing that each agent $i \in [2]$ satisfies γ-observability.

For any$i \in [2], h \in [2], b_1, b_2 \in \Delta(\mathcal{S})$, let

$$
\begin{aligned}
||\mathbb{O}_{i,h}^{\top}(b_1 - b_2)||_1 &= \sum_{o_{i,h}^1 \in [2]} \sum_{o_{i,h}^2 \in \mathcal{S}} |\sum_{s_h \in \mathcal{S}} (b_1(s_h) - b_2(s_h)) \mathbb{O}_{i,h}((o_{i,h}^1, o_{i,h}^2) \,|\, s_h)| \\
&\geq \sum_{o_{i,h}^2 \in \mathcal{S}} |\sum_{o_{i,h}^1 \in [2]} \sum_{s_h \in \mathcal{S}} (b_1(s_h) - b_2(s_h)) \mathbb{O}_{i,h}((o_{i,h}^1, o_{i,h}^2) \,|\, s_h)| \\
&= \sum_{o_{i,h}^2 \in \mathcal{S}} |\sum_{s_h \in \mathcal{S}} \sum_{o_{i,h}^1 \in [2]} (b_1(s_h) - b_2(s_h)) \mathbb{1}[o_{i,h}^1 = s_h^{i+2h-2}] (\frac{1-\alpha_2}{16} + \alpha_2 \mathbb{1}[o_{i,h}^2 = s_h])| \\
&= \sum_{o_{i,h}^2 \in \mathcal{S}} |\sum_{s_h \in \mathcal{S}} (b_1(s_h) - b_2(s_h)) (\frac{1-\alpha_2}{16} + \alpha_2 \mathbb{1}[o_{i,h}^2 = s_h])| \\
&= \sum_{o_{i,h}^2 \in \mathcal{S}} |\frac{1-\alpha_2}{16} (\sum_{s_h \in \mathcal{S}} (b_1(s_h) - b_2(s_h))) + \alpha_2 (b_1(o_{i,h}^2) - b_2(o_{i,h}^2))| \\
&= \sum_{o_{i,h}^2 \in \mathcal{S}} \alpha_2 |b_1(o_{i,h}^2) - b_2(o_{i,h}^2)| = \alpha_2 ||b_1 - b_2||_1.
\end{aligned}
$$

For any $i \in [2], h = 3, 4$, the proof is similar, where we replace $o_{i,h}^1 \in [2]$ with $o_{i,h}^1 \in \widetilde{Y}_i$ for $h = 3$ and replace the space $o_{i,h}^1 \in [2]$ with $\emptyset$ for $h = 4$.

- Problem $\mathcal{L}$ satisfies Assumption III.5, because the control actions $a_{1:4}$ do not influence underlying states, and we restrict the communication and control strategies do not use them as input. with $ct(1) = ct(2) = ct(3) = 1, ct(4) = 2$.

We will show as follows that computing a team-optimal strategy can give us a team-optimal strategy in $\mathcal{TD}$. Given $(g_{1:4}^{a,*}, g_{1:4}^{m,*})$ being a team optimal strategy of $\mathcal{L}$, firstly it will have no additional sharing at timesteps $h = 1, 2, 4$, namely, for $h = 1, 2, 4, \mathbb{P}(z_h^a \neq \{m_{1,h}, m_{2,h}\} \,|\, g_{1:4}^{a,*}, g_{1:4}^{m,*}) = 1$, since any additional sharing at timesteps $h = 1, 2, 4$ will incur a cost as high as 1, and cannot achieve the optimum. Also, for the additional sharing at timestep $h = 3$, agent i will definitely share $o_{i,3}$ and choose to share $o_{i,1}$ or $o_{i,2}$. Then $\forall \tau_{1,3^+} \in \mathcal{T}_{1,3^+}, g_{1,3}^{a,*}(\tau_{1,3^+}) = \begin{cases} o_{2,1} & \text{if } o_{2,1} \in \tau_{1,3^+} \\ o_{2,2} & \text{if } o_{2,2} \in \tau_{1,3^+} \end{cases}$ and $\forall \tau_{2,4^+} \in \mathcal{T}_{2,4^+}, g_{2,4}^{a,*}(\tau_{2,4^+}) = \begin{cases} o_{1,1} & \text{if } o_{1,1} \in \tau_{2,4^+} \\ o_{1,2} & \text{if } o_{1,2} \in \tau_{2,4^+} \end{cases}$, since such an action can achieve the optimal reward $r_3 = r_4 = 1$. Therefore, $J_{\mathcal{L}}(g_{1:H}^{a,*}, g_{1:H}^{m,*}) = \mathbb{E}[\sum_{h=1}^4 r_h - \kappa_h \,|\, g_{1:H}^{a,*}, g_{1:H}^{m,*}] = 2 - \mathbb{E}[\kappa_3 \,|\, g_{1:H}^{a,*}, g_{1:H}^{m,*}] = 2 - \mathbb{E}[\widetilde{c}(o_{1,3}^1, o_{2,3}^1, m_{1,3}, m_{2,3})]$, where $m_{1,3} = g_{1,3}^{m,*}(\{o_{1,1}, o_{1,2}, o_{1,3}\})$. Since κ_3 is independent of $\{o_{1,1}, o_{1,2}, o_{1,3}^2\}$, $\{o_{1,1}, o_{1,2}, o_{1,3}^2\}$ are useless information for agent 1 to choose $m_{1,3}$ and minimize the κ. Therefore, not using them in $g_{1,3}^{m,*}$ does not lose any optimality. Hence, we can consider the $g_{1,3}^{m,*}$ that only has $o_{1,3}^1$ as input. In the same way, we consider the $g_{2,3}^{m,*}$ that only has $o_{2,3}^1$ as input. Therefore, $J_{\mathcal{L}}(g_{1:H}^{a,*}, g_{1:H}^{m,*}) = 2 - \sum_{o_{1,3}^1, o_{1,3}^2, m_{1,3}, m_{2,3}} \frac{\widetilde{c}(o_{1,3}^1, o_{2,3}^1, m_{1,3}, m_{2,3})}{\alpha_1} g_{1,3}^{m,*}(m_{1,3} \,|\, o_{1,3}^1) g_{2,3}^{m,*}(m_{2,3} \,|\, o_{2,3}^1) \widetilde{p}(o_{1,3}^1, o_{2,3}^1)$. Then we can construct $\gamma_1 = g_{1,3}^{m,*}, \gamma_2 = g_{2,3}^{m,*}$, which minimize $\widetilde{J}$. Therefore, we can conclude that computing a team optimal strategy of $\mathcal{L}$ can give us a team optimal strategy of $\mathcal{TD}$. From the `NP-hardness` of the TDP problem [29], we complete our proof. □

3) Proof of Lemma III.6:

Proof of Lemma III.6. We prove this result by showing a reduction from the Team Decision problem. Given any TDP $\mathcal{TD} = (\widetilde{Y}_1, \widetilde{Y}_2, \widetilde{U}_1, \widetilde{U}_2, \widetilde{c}, \widetilde{p}, \widetilde{J})$ with $|\widetilde{U}_1| = |\widetilde{U}_2| = 2$, let $\widetilde{U}_1 = \{1, 2\}, \widetilde{U}_2 = \{1, 2\}$, then we can construct an $H = 5$ and 2-agent LTC $\mathcal{L}$ as follows:

- Underlying state: $\mathcal{S} = [2]^4$. For each $s_1 \in \mathcal{S}$, we can split s_1 into 4 parts as $s_1 = (s_1^1, s_1^2, s_1^3, s_1^4)$, where $s_1^1, s_1^2, s_1^3, s_1^4 \in [2]$. Similarly, $s_2, s_3, s_4, s_5 \in \mathcal{S}$ can be split in the same way.
- Initial state distribution: $\forall s_1 \in \mathcal{S}, \mu_1(s_1) = \frac{1}{16}$.
- Control action space: For $\forall i = 1, 2$, for $h = 1, 2$, $\mathcal{A}_{i,1} = \mathcal{A}_{i,2} = \{\emptyset\}$; For $h = 3$, $\mathcal{A}_{i,3} = \{(0, x), (x, 0) \,|\, x \in [2]\}$; We can write $a_{i,3} = (a_{i,3}^1, a_{i,3}^2), a_{i,3}^1, a_{i,3}^2 \in \{0, 1, 2\}$. For $h = 4, \mathcal{A}_{1,4} = [2], \mathcal{A}_{2,4} = \{\emptyset\}$; For $h = 5, \mathcal{A}_{2,5} = [2], \mathcal{A}_{1,5} = \{\emptyset\}$.
- Transition: $\forall s \in \mathcal{S}, a_1 \in \mathcal{A}_1, a_2 \in \mathcal{A}_2, a_3 \in \mathcal{A}_3, a_4 \in \mathcal{A}_4, a_5 \in \mathcal{A}_5, \mathbb{T}_1(s \,|\, s, a_1) = \mathbb{T}_2(s \,|\, s, a_2) = \mathbb{T}_3(s \,|\, s, a_3) = \mathbb{T}_4(s \,|\, s, a_4) = \mathbb{T}_5(s \,|\, s, a_5) = 1$. Note that under the transition dynamics above, $s_1 = s_2 = s_3 = s_4 = s_5$ always holds, for any $s_1 \in \mathcal{S}$.
- Observation space: $\mathcal{O}_{1,1} = \mathcal{O}_{2,1} = \mathcal{O}_{1,2} = \mathcal{O}_{2,2} = [2] \times \mathcal{S}$, $\mathcal{O}_{1,3} = \widetilde{Y_1} \times \mathcal{S}, \mathcal{O}_{2,3} = \widetilde{Y_2} \times \mathcal{S}, \mathcal{O}_{1,4} = \mathcal{O}_{2,4} = \mathcal{O}_{1,5} = \mathcal{O}_{2,5} = \mathcal{S}$; For each $i \in [2], h \in [2], o_{i,h} \in \mathcal{O}_{i,h}$, we can split $o_{i,h}$ into 2 parts as $o_{i,h} = (o_{i,h}^1, o_{i,h}^2)$, where $o_{i,h}^1 \in [2], o_{i,h}^2 \in \mathcal{S}$. For each $i \in [2], o_{i,3} \in \mathcal{O}_{i,3}$, similarly, we can split $o_{i,3}$ into 2 parts as $o_{i,3} = (o_{i,3}^1, o_{i,3}^2)$, where $o_{i,3}^1 \in \widetilde{Y}_i, o_{i,3}^2 \in \mathcal{S}$.

- The baseline sharing is null.
- Communication action space: For $i \in [2], h \in \{1,2,3,5\}, \mathcal{M}_{i,h} = \{0,1\}^{2h-1}$ and $\phi_{i,h}$ is defined as $\phi_{i,h}(p_{i,h^-}, m_{i,h}) = \{o_{i,k} \in p_{i,h^-} \mid k \le h, (2k-1)^{\text{th}} \text{ digit of } m_{i,h} \text{ is } 1\} \cup \{a_{i,k} \in p_{i,h^-} \mid k \le h-1, 2k^{\text{th}} \text{ digit of } m_{i,h} \text{ is } 1\} \cup \{m_{i,h}\}$; For $h = 4, \mathcal{M}_{i,4} = \{1,2\}, \phi_{i,h}(p_{i,h^-}, 1) = \{o_{i,3}, m_{i,h}\}, \phi_{i,h}(p_{i,h^-}, 2) = \{o_{i,3}, a_{i,3}, m_{i,h}\}$.
- Emission matrix: For any $i \in [2], h \in [2], s_h \in \mathcal{S}, o_{i,h} \in \mathcal{O}_{i,h}, \mathbb{O}_h(o_h \mid s_h) = \Pi_{i=1}^2 \mathbb{O}_{i,h}(o_{i,h} \mid s_h)$ and $\mathbb{O}_{i,h}(o_{i,h} \mid s_h)$ is defined as:

$$\mathbb{O}_{i,h}(o_{i,h} \mid s_h) = \begin{cases} \frac{1-\alpha_2}{16} & o_{i,h}^1 = s_h^{i+2h-2}, o_{i,h}^2 \neq s_h \\ \frac{1-\alpha_2}{16} + \alpha_2 & o_{i,h}^1 = s_h^{i+2h-2}, o_{i,h}^2 = s_h \\ 0 & \text{o.w.} \end{cases}.$$

For $i \in [2], s_3 \in \mathcal{S}, o_3 \in \mathcal{O}_3, \mathbb{O}_3(o_3 \mid s_3) = \mathbb{O}_3^1(o_3^1 \mid s_3)\mathbb{O}_3^2(o_3^2 \mid s_3), \mathbb{O}_3^2 = \Pi_{i=1}^2 \mathbb{O}_{i,3}^2(o_{i,3}^2 \mid s_3)$ is defined as:

$$\mathbb{O}_3^1(o_3^1 \mid s_3) = \widetilde{p}(o_{1,3}^1, o_{2,3}^1)$$
$$\mathbb{O}_{i,3}^2(o_3^2 \mid s_3) = \begin{cases} \frac{1-\alpha_2}{16} & o_{i,3}^2 \neq s_3 \\ \frac{1-\alpha_2}{16} + \alpha_2 & o_{i,3}^2 = s_3 \end{cases}.$$

And for $i \in [2], h = 4$ or $5, s_h \in \mathcal{S}, o_{i,h} \in \mathcal{O}_{i,h}, \mathbb{O}_h(o_h \mid s_h) = \Pi_{i=1}^2 \mathbb{O}_{i,h}(o_{i,h} \mid s_h)$ and $\mathbb{O}_{i,h}(o_{i,h} \mid s_h)$ is defined as:

$$\mathbb{O}_{i,h}(o_{i,h} \mid s_h) = \begin{cases} \frac{1-\alpha_2}{16} & o_{i,h} \neq s_h \\ \frac{1-\alpha_2}{16} + \alpha_2 & o_{i,h} = s_h \end{cases}.$$

- Reward functions:

$$\mathcal{R}_1(s_1, a_1) = \mathcal{R}_2(s_2, a_2) = \mathcal{R}_3(s_3, a_3) = 0, \quad \forall s_1, s_2, s_3 \in \mathcal{S}, a_1 \in \mathcal{A}_1, a_2 \in \mathcal{A}_2, a_3 \in \mathcal{A}_3;$$
$$\mathcal{R}_4(s_4, a_4) = \begin{cases} 1 & \text{if } a_{1,4} = s_4^2 \text{ or } a_{1,4} = s_4^4 \\ 0 & \text{o.w.} \end{cases};$$
$$\mathcal{R}_5(s_5, a_5) = \begin{cases} 1 & \text{if } a_{2,5} = s_5^1 \text{ or } a_{2,5} = s_5^3 \\ 0 & \text{o.w.} \end{cases}.$$

- Communication cost functions:

$$\forall h \in \{1,2,3,5\}, z_h^a \in \mathcal{Z}_h^a, \mathcal{K}_h(z_h^a) = 1 \text{ if } z_h^a \neq \{m_{1,h}, m_{2,h}\}, \text{ else } 0;$$
$$\mathcal{K}_4(z_4^a) = \begin{cases} \widetilde{c}(o_{1,3}^1, o_{2,3}^1, 1, 1)/\alpha_1 & \text{if } a_{1,3}, a_{2,3} \in z_3^a, a_{1,3}^1 = 0, a_{2,3}^1 = 0 \\ \widetilde{c}(o_{1,3}^1, o_{2,3}^1, 2, 1)/\alpha_1 & \text{if } a_{1,3}, a_{2,3} \in z_3^a, a_{1,3}^2 = 0, a_{2,3}^1 = 0 \\ \widetilde{c}(o_{1,3}^1, o_{2,3}^1, 1, 2)/\alpha_1 & \text{if } a_{1,3}, a_{2,3} \in z_3^a, a_{1,3}^1 = 0, a_{2,3}^2 = 0 \\ \widetilde{c}(o_{1,3}^1, o_{2,3}^1, 2, 2)/\alpha_1 & \text{if } a_{1,3}, a_{2,3} \in z_3^a, a_{1,3}^2 = 0, a_{2,3}^2 = 0 \\ 1 & \text{o.w.} \end{cases}.$$

Let $\alpha_0 = \max_{y_1, y_2, u_1, u_2} \widetilde{c}(y_1, y_2, u_1, u_2)$, set $\alpha_1 = 2\alpha_0$, and restrict agents to decide their communication strategies only based on their common information. Under such a construction, $\mathcal{L}$ satisfies the following conditions:

- Problem $\mathcal{L}$ is QC: For $\forall i_1, i_2 \in [2], h_1, h_2 \in [4]$, agent (i_1, h_1) does not influence (i_2, h_2) because agent (i_1, h_1) cannot influence the observation of agent (i_2, h_2), and the baseline sharing is null.
- Problem $\mathcal{L}$ satisfies Assumptions III.1 and III.7: We prove this by showing that each agent $i \in [2]$ satisfies γ-observability.

For $\forall i \in [2], h \in [2], b_1, b_2 \in \Delta(\mathcal{S})$, let

$$
\begin{aligned}
||\mathbb{O}_{i,h}^{\top}(b_1 - b_2)||_1 &= \sum_{o_{i,h}^1 \in [2]} \sum_{o_{i,h}^2 \in \mathcal{S}} |\sum_{s_h \in \mathcal{S}} (b_1(s_h) - b_2(s_h)) \mathbb{O}_{i,h}((o_{i,h}^1, o_{i,h}^2) \,|\, s_h)| \\
&\geq \sum_{o_{i,h}^2 \in \mathcal{S}} |\sum_{o_{i,h}^1 \in [2]} \sum_{s_h \in \mathcal{S}} (b_1(s_h) - b_2(s_h)) \mathbb{O}_{i,h}((o_{i,h}^1, o_{i,h}^2) \,|\, s_h)| \\
&= \sum_{o_{i,h}^2 \in \mathcal{S}} |\sum_{s_h \in \mathcal{S}} \sum_{o_{i,h}^1 \in [2]} (b_1(s_h) - b_2(s_h)) \mathbb{1}[o_{i,h}^1 = s_h^{i+2h-2}] (\frac{1-\alpha_2}{16} + \alpha_2 \mathbb{1}[o_{i,h}^2 = s_h])| \\
&= \sum_{o_{i,h}^2 \in \mathcal{S}} |\sum_{s_h \in \mathcal{S}} (b_1(s_h) - b_2(s_h)) (\frac{1-\alpha_2}{16} + \alpha_2 \mathbb{1}[o_{i,h}^2 = s_h])| \\
&= \sum_{o_{i,h}^2 \in \mathcal{S}} |\frac{1-\alpha_2}{16} (\sum_{s_h \in \mathcal{S}} (b_1(s_h) - b_2(s_h))) + \alpha_2 (b_1(o_{i,h}^2) - b_2(o_{i,h}^2))| \\
&= \sum_{o_{i,h}^2 \in \mathcal{S}} \alpha_2 |b_1(o_{i,h}^2) - b_2(o_{i,h}^2)| = \alpha_2 ||b_1 - b_2||_1.
\end{aligned}
$$

For $\forall i \in [2], h = 3, 4$, the proof is similar, by replacing $o_{i,h}^1 \in [2]$ with $o_{i,h}^1 \in \widetilde{Y}_i$ for $h = 3$ and replacing the space $o_{i,h}^1 \in [2]$ with $\{\emptyset\}$ for $h = 4, 5$.

- Problem $\mathcal{L}$ satisfies Assumption III.4 since we restrict agents to decide their communication strategies only based on the common information, with $ct(1) = ct(2) = ct(3) = ct(4) = 1, ct(5) = 2$.

Now, we show that any team-optimal strategy of $\mathcal{L}$ will give us the decision rules γ_1, γ_2 to solve $\mathcal{TD}$. Let $(g_{1:5}^{a,*}, g_{1:5}^{m,*})$ be a team-optimal strategy. First, $\forall \tau_{i,4^-} \in \mathcal{T}_{i,4^-}, g_{i,4}^{m,*}(\tau_{i,4^-}) = 2$, since otherwise it will have communication cost $\kappa_{i,3} = 1$, and cannot achieve the team optimum. Define $\overline{g}_{1:5}^a, \overline{g}_{1:5}^m$ as

$$
\begin{aligned}
&\forall \tau_{i,3^+} \in \mathcal{T}_{i,3^+}, \overline{g}_{i,3^+}^a(\tau_{i,3^+}) = \begin{cases} (o_{i,1}^1, 0) & \text{if } a_{i,3} = g_{i,3^+}^{a,*}(\tau_{i,3^+}), a_{i,3}^1 = 0 \\ (0, o_{i,2}^1) & \text{o.w.} \end{cases} \\
&\forall \tau_{1,4^+} \in \mathcal{T}_{1,4^+}, \overline{g}_{1,4^+}^a(\tau_{1,4^+}) = \begin{cases} a_{2,4}^1 & \text{if } a_{2,4}^1 \neq 0 \\ a_{2,4}^2 & \text{o.w.} \end{cases} \\
&\overline{g}_{1:5}^m = g_{1:5}^{m,*}, \overline{g}_{1:2}^a = g_{1:2}^{a,*}, \overline{g}_{4:5}^a = g_{4:5}^{a,*}.
\end{aligned}
$$

Then, $J_{\mathcal{L}}(\overline{g}_{1:5}^a, \overline{g}_{1:5}^m) - J_{\mathcal{L}}(g_{1:5}^{a,*}, g_{1:5}^{m,*}) \geq 0$. Hence $(\overline{g}_{1:5}^a, \overline{g}_{1:5}^m)$ is a team-optimal strategy. Then, $J_{\mathcal{L}}(\overline{g}_{1:5}^a, \overline{g}_{1:5}^m) = 2 - \mathbb{E}[\kappa_4 \,|\, \overline{g}_{1:5}^a, \overline{g}_{1:5}^m] = 2 - \mathbb{E}[\kappa_4 \,|\, \overline{g}_3^a]$, where $\overline{g}_3^a$ minimizes κ_4. Note that $\tau_{i,3^+} = \{o_{i,1}, o_{i,2}, o_{i,3}\}$. Since κ_4 is independent of $o_{i,1}, o_{i,2}, o_{i,3}^2$, they are useless information for agent i to choose $a_{i,3}$ and minimize κ_4. Therefore, only using $o_{i,3}^1$ to determine $a_{i,3}$ does not lose any optimality, and we can consider $g_{1,3}^{a,*}$ that has only $o_{i,3}^1$ as input. In the same way, we consider $g_{2,3}^{a,*}$ that has only $o_{i,3}^1$ as input. Then, we can construct $\gamma_1 = g_{1,3}^{a,*}, \gamma_2 = g_{2,3}^{a,*}$ as decision rules that minimize $\widetilde{J}$. Therefore, we can conclude that computing a team-optimal strategy of $\mathcal{L}$ can give us a team-optimal strategy of $\mathcal{TD}$. From the `NP-hardness` of the TDP problem [29], we complete our proof. □

4) Proof of Lemma III.8:

Proof. We prove this by showing a reduction from the hardness of finding an ϵ-optimal strategy in POMDPs. Given any POMDP $\mathcal{P} = (\mathcal{S}^{\mathcal{P}}, \mathcal{A}^{\mathcal{P}}, \mathcal{O}^{\mathcal{P}}, \{\mathbb{O}_h^{\mathcal{P}}\}_{h \in [H^{\mathcal{P}}]}, \{\mathbb{T}_h^{\mathcal{P}}\}_{h \in [H^{\mathcal{P}}]}, \{\mathcal{R}_h^{\mathcal{P}}\}_{h \in [H^{\mathcal{P}}]}, \mu_1^{\mathcal{P}})$, we can construct a LTC $\mathcal{L}$ with 2 agents as follows:

- Number of agents: $n = 2$; length of episode: $H = H^{\mathcal{P}}$.
- $\mathcal{S} = \mathcal{S}^{\mathcal{P}}, \forall s \in \mathcal{S}$.
- Initial state distribution: $\forall s_1 \in \mathcal{S}, \mu_1(s_1) = \mu_1^{\mathcal{P}}(s_1)$.
- Control action space: $\forall h \in [H]$, $\mathcal{A}_{1,h} = \mathcal{A}_h^{\mathcal{P}}, \mathcal{A}_{2,h} = \{\emptyset\}$.
- Transition: $\forall s_h, s_{h+1} \in \mathcal{S}, a_h \in \mathcal{A}_h, \mathbb{T}_h(s_{h+1} \,|\, s_h, a_h) = \mathbb{T}_h^{\mathcal{P}}(s_{h+1} \,|\, s_h, a_{1,h})$.
- Observation space: $\forall h \in [H], \mathcal{O}_{1,h} = \mathcal{O}^{\mathcal{P}}, \mathcal{O}_{2,h} = \mathcal{S}$.
- Emission matrix: For any $h \in [H], \forall o_h \in \mathcal{O}_h, s_h \in \mathcal{S}$, $\mathbb{O}_h(o_h \,|\, s_h) = \mathbb{O}_h^{\mathcal{P}}(o_{1,h} \,|\, s_h) \mathbb{1}[o_{2,h} = s_h]$.
- Reward functions: For any $h \in [H], i \in [2], s_h \in \mathcal{S}, a_h \in \mathcal{A}_h, \mathcal{R}_h(s_h, a_h) = \mathcal{R}^{\mathcal{P}}(s_h, a_{1,h})/H$.
- The baseline sharing: For any $h \in [H], z_h^b = \{o_{1,h}, a_{1,h-1}\}$.
- Communication action space: For any $h \in [H], \mathcal{M}_{1,h} = \{\emptyset\}, \mathcal{M}_{2,h} = \{0,1\}^h$. For any $p_{1,h^-} \in \mathcal{P}_{1,h^-}, p_{2,h^-} \in \mathcal{P}_{2,h^-}, m_h \in \mathcal{M}_h, \phi_{1,h}(p_{1,h^-}, m_{1,h}) = \{m_{1,h}\}, \phi_{2,h}(p_{2,h^-}, m_{2,h}) = \{o_{2,k} \,|\, k\text{-th digit of } p_{2,h^-} \text{ is 1 and } o_{2,k} \in p_{i,h^-}\} \cup \{m_{2,h}\}$.

- Communication cost functions: For any $h \in [H], z_h^a \in \mathcal{Z}_h^a, \mathcal{K}_h(z_h^a) = \mathbb{1}[z_h^a \neq \{m_h\}]$. It means the communication cost is 1 unless there is no additional sharing.
- We restrict that the communication strategy can only use c_h as input, and remove $a_{2,t}$ in τ_h for any $h > t$.

We first verify that $\mathcal{L}$ is QC and satisfies Assumptions III.1, III.4, III.5, and IV.7.

- $\mathcal{L}$ is QC: For any $\forall h_1 < h_2 \leq H$, agent $(2, h_1)$ does not influence agent $(1, h_2)$ under baseline sharing since agent $(2, h_1)$ does not influence $s_h^1, \forall h \in [H]$, then does not influence $o_{1,h}, \forall h \in [H]$, and thus not influencing agent $(1, h_1)$. For any $\forall h_1 < h_2 \leq H$, under baseline sharing, $p_{1,h^-} = \emptyset$. Then $\sigma(\tau_{1,h_1^-}) \subseteq \sigma(c_{h_1^-}) \subseteq \sigma(c_{h_2^-}) \subseteq \sigma(\tau_{2,h_2^-})$.
- $\mathcal{L}$ satisfies Assumption III.1: For any $h \in [H], b_1, b_2 \in \Delta(\mathcal{S})$, $\mathbb{O}_h$ satisfies

$$
\begin{aligned}
||\mathbb{O}_h^\top (b_1 - b_2)||_1 &= \sum_{o_{1,h} \in \mathcal{O}^{\mathcal{P}}} \sum_{o_{2,h} \in \mathcal{S}} |\sum_{s_h \in \mathcal{S}} (b_1(s_h) - b_2(s_h)) \mathbb{O}_h((o_{1,h}, o_{2,h}) \,|\, s_h)| \\
&\geq \sum_{o_{2,h} \in \mathcal{S}} |\sum_{o_{1,h} \in \mathcal{O}^{\mathcal{P}}} \sum_{s_h \in \mathcal{S}} (b_1(s_h) - b_2(s_h)) \mathbb{O}_{1,h}(o_{1,h} \,|\, s_h) \mathbb{O}_{2,h}(o_{2,h} \,|\, s_h)| \\
&= \sum_{o_{2,h} \in \mathcal{S}} |\sum_{s_h \in \mathcal{S}} (b_1(s_h) - b_2(s_h)) \mathbb{O}_{2,h}(o_{2,h} \,|\, s_h) \sum_{o_{1,h} \in \mathcal{O}^{\mathcal{P}}} \mathbb{O}_{1,h}(o_{1,h} \,|\, s_h)| \\
&= \sum_{o_{2,h} \in \mathcal{S}} |\sum_{s_h \in \mathcal{S}} (b_1(s_h) - b_2(s_h)) \mathbb{1}[o_{2,h} = s_h] \\
&= \sum_{o_{2,h} \in \mathcal{S}} |b_1(o_{2,h}) - b_2(o_{2,h})| = ||b_1 - b_2||_1.
\end{aligned}
$$

- $\mathcal{L}$ satisfies Assumption III.4: For any $h \in [H]$, we restrict that each agent decides $m_{i,h}$ based on c_h.
- $\mathcal{L}$ satisfies Assumption III.5: For any $h \in [H], a_{2,h}$ does not influence s_{h+1}, and it is removed from τ.

Agent 2 will share nothing through additional sharing, otherwise it will suffer the communication cost $\kappa_h = 1 > \max \sum_{h=1}^H \mathcal{R}_h(s_h, a_h)$ and cannot achieve optimum. Hence, agent 2 is the dummy player. Therefore, any $(g_{1:H}^{a,*}, g_{1:H}^{m,*})$ be an ϵ/H-team optimal strategy of $\mathcal{L}$ will directly gives the ϵ-optimal of $\mathcal{P}$ as $\{g_{1,1:H}^{a,*}\}_{h \in [H]}$. From Proposition .5, we can complete our proof. □

C. Deferred Details of §IV

1) Proof of Theorem IV.2:

Proof. We prove the following lemma first.

Lemma .7. Let $\mathcal{L}$ be a QC LTC problem satisfying Assumptions III.5 and III.7, and $\mathcal{D}_{\mathcal{L}}$ be the reformulated Dec-POMDP. Then for any $i_1, i_2 \in [n], t_1, t_2 \in [H]$, if agent $(i_1, 2t_1)$ influences agent $(i_2, 2t_2)$ in $\mathcal{D}_{\mathcal{L}}$, then $\sigma(\tau_{i_1,t_1^-}) \subseteq \sigma(\tau_{i_2,t_2^-})$ in $\mathcal{L}$. Moreover, if $\mathcal{L}$ is sQC, then $\sigma(a_{i_1,t_1}) \subseteq \sigma(\tau_{i_2,t_2^-})$.

Proof. We prove this case-by-case as follows:

- If a_{i_1,t_1} influences the underlying state s_{t_1+1}, then from Assumption III.7, agent (i_1, t_1) influences o_{-i_1,t_1+1}, so there must exist $i_3 \neq i_1$, such that agent (i_1, t_1) influences o_{i_3,t_1+1}. From part (e) of Assumption II.1 and $t_1 < t_2$, we know $o_{i_3,t_1+1} \in \tau_{i_3,(t_1+1)^-} \subseteq \tau_{i_3,t_2^-}$ even under no additional sharing, and then we get agent (i_1, t_1) influences agent (i_3, t_2) in $\overline{\mathcal{D}}_{\mathcal{L}}$ (the Dec-POMDP induced by $\mathcal{L}$). From Lemma .4, it holds that $\sigma(\tau_{i_1,t_1^-}) \subseteq \sigma(\tau_{i_3,t_2^-})$. From Assumption II.2 and $i_3 \neq i_1$, we know $\sigma(\tau_{i_1,t_1^-}) \subseteq \sigma(c_{t_2^-}) \subseteq \sigma(\tau_{i_2,t_2^-})$. Similarly, by Lemma .4, if $\mathcal{L}$ is sQC, we have $\sigma(a_{i_1,t_1}) \subseteq \sigma(\tau_{i_3,t_2^-})$ from Assumption II.2, and $\sigma(a_{i_1,t_1}) \subseteq \sigma(c_{t_2^-}) \subseteq \sigma(\tau_{i_2,t_2^-})$ from Assumption II.2.
- If a_{i_1,t_1} does not influence s_{t_1+1}, from Assumption III.5, $\forall t > t_1, a_{i_1,t_1} \notin \tau_{t^-}$ and $a_{i_1,t_1} \notin \tau_{t^+}$. Then in $\mathcal{D}_{\mathcal{L}}$, agent $(i_1, 2t_1)$ does not influence s_{2t_1+1} and o_{2t_1+1}. Thus it does not influence $\widetilde{\tau}_{i,2t_1+1}, \forall i \in [n]$, and then it does not influence z_{2t_1+1} and $\widetilde{a}_{i,2t_1+1}, \forall i \in [n]$. Thus, it does not influence $\widetilde{z}_{2t_1+2}$, and further does not influence $\widetilde{\tau}_{i,2t_1+2}$ and $\widetilde{a}_{i,2t_1+2}, \forall i \in [n]$, either. From induction, we know agent $(i_1, 2t_1)$ does not influence agent $(i_2, 2t_2)$, which leads to a contradiction to the premise of the lemma.

This completes the proof. □

We now go back to proving the theorem. Firstly, we prove the QC cases. To show $\mathcal{D}_{\mathcal{L}}$ is QC, we need to prove $\forall i_1, i_2 \in [n], h_1, h_2 \in [\widetilde{H}]$, if agent (i_1, h_1) influences agent (i_2, h_2) with $h_1 < h_2$, then $\sigma(\widetilde{\tau}_{i_1,h_1}) \subseteq \sigma(\widetilde{\tau}_{i_2,h_2})$, where we use $\widetilde{\tau}_{i,h}$ to denote the available information of agent (i, h) in $\mathcal{D}_{\mathcal{L}}$. We prove this by considering the following cases:

1) If $h_1 = 2t_1 - 1$ with $t_1 \in [H]$, by the construction of $\mathcal{D}_{\mathcal{L}}$ and Assumption III.4, we have $\widetilde{\tau}_{i_1,h_1} = \widetilde{c}_{h_1} = c_{t_1^-} \subseteq \widetilde{\tau}_{i_2,h_2}$, since common information accumulates over time by definition, and will always be included in the available information $\widetilde{\tau}_{i,h}$ in later steps. Thus, $\sigma(\widetilde{\tau}_{i_1,h_1}) \subseteq \sigma(\widetilde{\tau}_{i_2,h_2})$.

2) If $h_1 = 2t_1, h_2 = 2t_2$ with $t_1, t_2 \in [H]$, then $\widetilde{\tau}_{i_1,h_1} = \tau_{i_1,t_1^+} = \tau_{i_1,t_1^-} \cup z^a_{t_1}$ by definition. Consider agent (i_1, t_1) and (i_2, t_2) in $\mathcal{L}$. From Lemma .7, we know $\sigma(\tau_{i_1,t_1^-}) \subseteq \sigma(\tau_{i_2,t_2^-}) \subseteq \sigma(\tau_{i_2,t_2^+})$. Also, $z^a_{t_1} \subseteq c_{t_1^+} \subseteq c_{t_2^+} \subseteq \tau_{i_2,t_2^+} = \widetilde{\tau}_{i_2,h_2}$ by the accumulation of c_{h^+} over time. Thus, we have $\sigma(\widetilde{\tau}_{i_1,h_1}) \subseteq \sigma(\widetilde{\tau}_{i_2,h_2})$.
3) If $h_1 = 2t_1, h_2 = 2t_2 - 1, t_1, t_2 \in [H]$, then $\widetilde{\tau}_{i_2,h_2} = \widetilde{c}_{h_2}$, then $\exists i_3 \in [n], i_3 \neq i_1, \widetilde{\tau}_{i_2,h_2} \subseteq \widetilde{c}_{h_2+1} \subseteq \widetilde{\tau}_{i_3,h_2+1}$. From agent (i_1, h_1) influences (i_2, h_2), we know agent (i_1, h_1) also influences agent $(i_3, h_2 + 1)$ in $\mathcal{D}_{\mathcal{L}}$, hence agent (i_1, t_1) influences agent (i_2, t_2) in $\mathcal{L}$. Since $\mathcal{L}$ is QC, we know $\sigma(\tau_{i_1,t_1^-}) \subseteq \sigma(\tau_{i_3,t_2^-})$. From Assumption II.2 and $i_1 \neq i_3$, we know $\sigma(\widetilde{\tau}_{i_1,h_1}) = \sigma(\tau_{i_1,t_1^-}) \subseteq \sigma(c_{t_2^-}) = \sigma(\widetilde{\tau}_{i_2,h_2})$.

Second, we prove the sQC case. In $\mathcal{D}_{\mathcal{L}}$, for any $i_1, i_2 \in [n], h_1, h_2 \in [\widetilde{H}]$, agent (i_1, h_1) influences (i_2, h_2). From the proof above, we know $\sigma(\widetilde{\tau}_{i_1,h_1}) \subseteq \sigma(\widetilde{\tau}_{i_2,h_2})$. We only need to prove $\sigma(\widetilde{a}_{i_1,h_1}) \subseteq \sigma(\widetilde{\tau}_{i_2,h_2})$.

1) If $h_1 = 2t_1 - 1$ with $t_1 \in [H]$, then we know $\widetilde{a}_{i_1,h_1} = m_{i_1,t}$. From Assumption II.1, we know that $m_{i_1,t} \subseteq z^a_{i_1,t}$. Then we get $\sigma(\widetilde{a}_{i_1,h_1}) \subseteq \sigma(\widetilde{z}_{i_1,h_1+1}) \subseteq \sigma(\widetilde{c}_{h_2}) \subseteq \sigma(\widetilde{\tau}_{i_2,h_2})$.
2) If $h_1 = 2t_1, h_2 = 2t_2$ with $t_1, t_2 \in [H]$, then from Lemma .7, we know that $\sigma(\widetilde{a}_{i_1,h_1}) \subseteq \sigma(\widetilde{\tau}_{i_2,h_2})$.
3) If $h_1 = 2t_1, h_2 = 2t_2 - 1, t_1, t_2 \in [H]$, then $\widetilde{\tau}_{i_2,h_2} = \widetilde{c}_{h_2}$, then $\exists i_3 \in [n], i_3 \neq i_1, \widetilde{\tau}_{i_2,h_2} \subseteq \widetilde{c}_{h_2+1} \subseteq \widetilde{\tau}_{i_3,h_2+1}$. From agent (i_1, h_1) influences (i_2, h_2), we know agent (i_1, h_1) also influences agent $(i_3, h_2 + 1)$ in $\mathcal{D}_{\mathcal{L}}$, hence agent (i_1, t_1) influences agent (i_2, t_2) in $\mathcal{L}$. Since $\mathcal{L}$ is sQC, we know $\sigma(a_{i_1,t_1^-}) \subseteq \sigma(\tau_{i_3,t_2^-})$. From Assumption II.2 and $i_1 \neq i_3$, we know $\sigma(\widetilde{a}_{i_1,h_1}) = \sigma(a_{i_1,t_1}) \subseteq \sigma(c_{t_2^-}) = \sigma(\widetilde{\tau}_{i_2,h_2})$.

This completes the proof. □

2) Proof of Lemma IV.3:

Proof. From the construction of $\mathcal{D}^\dagger_{\mathcal{L}}$, since $\mathcal{D}^\dagger_{\mathcal{L}}$ requires agent to share more than $\mathcal{D}_{\mathcal{L}}$, it is easy to observe the fact that $\forall h \in [\widetilde{H}], i \in [n], \widetilde{c}_h \subseteq \breve{c}_h, \widetilde{\tau}_{i,h} \subseteq \breve{\tau}_{i,h}$.

Let $i_1, i_2 \in [n], h_1, h_2 \in [\widetilde{H}], h_1 < h_2$, and agent (i_1, h_1) influences agent (i_2, h_2) in $\mathcal{D}^\dagger_{\mathcal{L}}$.

- If $h_1 = 2t_1 - 1$ with $t_1 \in [H]$, then h_1 is communication step. So $\breve{\tau}_{i_1,h_1} = \breve{c}_{h_1} \subseteq \breve{c}_{h_2}$, and $\widetilde{a}_{i_1,h_1} = m_{i_1,t_1} \subseteq \breve{c}_{h_1+1} \subseteq \breve{c}_{h_2}$ from Assumption II.1. Therefore, we have $\sigma(\breve{\tau}_{i_1,h_1}) \cup \sigma(\breve{a}_{i_1,h_1}) \subseteq \sigma(\breve{c}_{h_1}) \subseteq \sigma(\breve{\tau}_{i_2,h_2})$.
- If $h_1 = 2t_1, h_2 = 2t_2 - 1$ with $t_1, t_2 \in [H]$, then $\breve{\tau}_{i_2,h_2} = \breve{c}_{h_2}$. If agent (i_1, h_1) does not influence (i_2, h_2) in $\mathcal{D}_{\mathcal{L}}$, but agent (i_1, h_1) influences (i_2, h_2) in $\mathcal{D}^\dagger_{\mathcal{L}}$, then it means $\breve{a}_{i_1,h_1} \in \breve{\tau}_{i_2,h_2}$ but $\widetilde{a}_{i_1,h_1} \notin \widetilde{\tau}_{i_2,h_2}$. This can only happen when $\sigma(\widetilde{\tau}_{i_1,h_1}) \subseteq \sigma(\widetilde{c}_{h_2}) \subseteq \sigma(\breve{c}_{h_2})$, and $\widetilde{a}_{i_1,h_1} \subseteq \breve{c}_{h_2}$. Also, from the construction of $\mathcal{D}^\dagger_{\mathcal{L}}$, we know that $\breve{\tau}_{i_1,h_1} \backslash \widetilde{\tau}_{i_1,h_1} \subseteq \breve{c}_{h_1}$. Therefore, we have $\sigma(\breve{\tau}_{i_1,h_1}) \cup \sigma(\widetilde{a}_{i_1,h_1}) \subseteq \sigma(\breve{c}_{h_2}) \subseteq \sigma(\breve{\tau}_{i_2,h_2})$.
  If agent (i_1, h_1) influences (i_2, h_2) in $\mathcal{D}_{\mathcal{L}}$, then from QC of $\mathcal{D}_{\mathcal{L}}$, we know that $\sigma(\widetilde{\tau}_{i_1,h_1}) \subseteq \sigma(\widetilde{c}_{h_2})$, then from the construction of $\mathcal{D}^\dagger_{\mathcal{L}}$, we have $\widetilde{a}_{i_1,h_1} \in \breve{c}_{h_2}$. Still, we have $\breve{\tau}_{i_1,h_1} \backslash \widetilde{\tau}_{i_1,h_1} \subseteq \breve{c}_{h_1}$. Therefore, $\sigma(\breve{\tau}_{i_1,h_1}) \cup \sigma(\widetilde{a}_{i_1,h_1}) \subseteq \sigma(\breve{\tau}_{i_2,h_2})$.
- If $h_1 = 2t_1, h_2 = 2t_2$ with $t_1, t_2 \in [H]$. If agent (i_1, h_1) does not influence (i_2, h_2) in $\mathcal{D}_{\mathcal{L}}$, then it means sharing $\widetilde{a}_{i_1,h_1}$ leads to the influence. Then, $\sigma(\widetilde{\tau}_{i_1,h_1}) \subseteq \sigma(\widetilde{c}_{h_2}) \subseteq \sigma(\breve{c}_{h_2})$, and $\widetilde{a}_{i_1,h_1} \subseteq \breve{c}_{h_2}$. We can conclude $\sigma(\breve{\tau}_{i_1,h_1}) \cup \sigma(\widetilde{a}_{i_1,h_1}) \subseteq \sigma(\breve{c}_{h_2}) \subseteq \sigma(\breve{\tau}_{i_2,h_2})$.
  Now we consider the case that agent (i_1, h_1) influences (i_2, h_2) in $\mathcal{D}_{\mathcal{L}}$. If $i_1 \neq i_2$, then we have $\widetilde{\tau}_{i_1,h_1} \subseteq \widetilde{\tau}_{i_2,h_2}$. From Assumption II.2, and $i_1 \neq i_2$, we know $\widetilde{\tau}_{i_1,h_1} \subseteq \widetilde{c}_{h_2}$. Then, from the construction of $\mathcal{D}^\dagger_{\mathcal{L}}$, we have $\widetilde{a}_{i_1,h_1} \subseteq \breve{c}_{h_2}$. Finally, we have $\sigma(\breve{\tau}_{i_1,h_1}) \cup \sigma(\widetilde{a}_{i_1,h_1}) \subseteq \sigma(\breve{\tau}_{i_2,h_2})$.
  If $i_1 = i_2$, then from the perfect recall of $\mathcal{L}$, we know that $\widetilde{\tau}_{i_1,h_1} \cup \widetilde{a}_{i_1,h_1} \subseteq \widetilde{\tau}_{i_2,h_2}$. From $\breve{\tau}_{i_1,h_1} \backslash \widetilde{\tau}_{i_1,h_1} \subseteq \breve{c}_{h_1}$, we conclude $\sigma(\breve{\tau}_{i_1,h_1}) \cup \sigma(\widetilde{a}_{i_1,h_1}) \subseteq \sigma(\breve{\tau}_{i_2,h_2})$.

This completes the proof. □

3) Proof of Theorem IV.4:

Proof. We claim that given any strategy $\breve{g}_{1:\breve{H}}$ and $\widetilde{g}_{1:\widetilde{H}} = \varphi(\breve{g}_{1:\breve{H}}, \mathcal{D}_{\mathcal{L}})$, $J_{\mathcal{D}^\dagger_{\mathcal{L}}}(\breve{g}_{1:\breve{H}}) = J_{\mathcal{D}_{\mathcal{L}}}(\widetilde{g}_{1:\widetilde{H}})$, where the function φ is given by Algorithm 3. In order to prove this statement, we prove that $\widetilde{g}_{i,h}(\widetilde{\tau}_{i,h}) = \breve{g}_{i,h}(\breve{\tau}_{i,h})$ always holds for any $\widetilde{\tau}_{i,h}$, which is equivalent to prove that for any $i \in [n], h \in [\widetilde{H}]$, and $\widetilde{\tau}_{i,h}$, Algorithm 3 can compute the associated $\breve{\tau}_{i,h}$ from the expansion in Eq. (IV.2), and use it as the input of $\breve{g}_{i,h}$ (Line 12). Let $\breve{\tau}'_{i,h}$ be the input of $\breve{g}_{i,h}$ used in Algorithm 3, then it holds that $\widetilde{\tau}_{i,h} \subseteq \breve{\tau}'_{i,h}$. We now aim to show that $\breve{\tau}_{i,h} = \breve{\tau}'_{i,h}$:

- For any $j \in [n], t < h$ such that $\widetilde{a}_{j,t} \in \breve{\tau}_{i,h} \backslash \widetilde{\tau}_{i,h}$, then it must hold that $\sigma(\widetilde{\tau}_{j,t}) \subseteq \sigma(\widetilde{c}_h)$ from Eq. (IV.2). From Algorithm 3, we know that $\widetilde{a}_{j,t} \in \breve{\tau}'_{i,h}$, and thus $\breve{\tau}_{i,h} \subseteq \breve{\tau}'_{i,h}$.
- For any $j \in [n], t < h$ such that $\widetilde{a}_{j,t} \in \breve{\tau}'_{i,h} \backslash \widetilde{\tau}_{i,h}$, then by Algorithm 3, it must hold that $\sigma(\widetilde{\tau}_{j,t}) \subseteq \sigma(\widetilde{c}_h)$. Therefore, from Eq. (IV.2), we know that $\widetilde{a}_{j,t} \in \breve{\tau}_{i,h}$, and thus $\breve{\tau}'_{i,h} \subseteq \breve{\tau}_{i,h}$.

Therefore, we conclude that $\breve{\tau}_{i,h} = \breve{\tau}'_{i,h}$ and prove the statement.

Since $\mathcal{D}^\dagger_{\mathcal{L}}$ has larger strategy spaces, i.e., $\max_{\widetilde{g}_{1:\widetilde{H}} \in \widetilde{G}_{1:\widetilde{H}}} J_{\mathcal{D}_{\mathcal{L}}}(\widetilde{g}_{1:\widetilde{H}}) \leq \max_{\breve{g}_{1:\breve{H}} \in \breve{G}_{1:\breve{H}}} J_{\mathcal{D}^\dagger_{\mathcal{L}}}(\breve{g}_{1:\breve{H}})$. Let $\breve{g}^*_{1:\breve{H}}$ be the strategy satisfying $J_{\mathcal{D}^\dagger_{\mathcal{L}}}(\breve{g}^*_{1:\breve{H}}) \geq \max_{\breve{g}_{1:\breve{H}} \in \breve{G}_{1:\breve{H}}} J_{\mathcal{D}^\dagger_{\mathcal{L}}}(\breve{g}_{1:\breve{H}}) - \epsilon$. Then, we have $J_{\mathcal{D}_{\mathcal{L}}}(\varphi(\breve{g}^*_{1:\breve{H}}, \mathcal{D}_{\mathcal{L}})) = J_{\mathcal{D}^\dagger_{\mathcal{L}}}(\breve{g}^*_{1:\breve{H}}) \geq \max_{\breve{g}_{1:\breve{H}} \in \breve{G}_{1:\breve{H}}} J_{\mathcal{D}^\dagger_{\mathcal{L}}}(\breve{g}_{1:\breve{H}}) - \epsilon \geq \max_{\widetilde{g}_{1:\widetilde{H}} \in \widetilde{G}_{1:\widetilde{H}}} J_{\mathcal{D}_{\mathcal{L}}}(\widetilde{g}_{1:\widetilde{H}}) - \epsilon$. Thus, $\varphi(\breve{g}^*_{1:\breve{H}}, \mathcal{D}_{\mathcal{L}})$ is an ϵ-team optimal strategy of $\mathcal{D}_{\mathcal{L}}$. □

Lemma .8. For any given strategy $\breve{g}_{1:\breve{H}} \in \breve{\mathcal{G}}_{1:\breve{H}}$, implementing Algorithm 4 in $\mathcal{D}_{\mathcal{L}}$ is equivalent to implementing $\varphi(\breve{g}_{1:\breve{H}}, \mathcal{D}_{\mathcal{L}})$ in $\mathcal{D}_{\mathcal{L}}$.

Proof. As shown in Theorem IV.4, Algorithm 3 can compute the associated $\breve{\tau}_{i,h}$ and use it as the input of $\breve{g}_{i,h}$ (Line 12). Therefore, it suffices to prove that Algorithm 4 can also compute the associated $\breve{\tau}_{i,h}$ and use it as the input of $\breve{g}_{i,h}$ (Line 6), i.e., Algorithm 5 can output the associated $\breve{\tau}_{i,h}$ from $\widetilde{\tau}_{i,h}$ and $\breve{g}_{1:h-1}$. We prove this by induction.
Firstly, when $h = 1$, it holds for any $i \in [n]$ such that $\widetilde{\tau}_{i,1} = \breve{\tau}_{i,1}$. In Algorithm 5, when $h = 1$, it will never enter the for loop, and thus the output is $\widetilde{\tau}_{i,1} = \breve{\tau}_{i,1}$.
Secondly, we assume for any $h < t$, the hypothesis holds. Then for $h = t$, given any $\widetilde{\tau}_{i,t} \in \widetilde{\mathcal{T}}_{i,t}$ and $\breve{g}_{1:t-1}$, let $\breve{\tau}'_{i,t}$ be the output of Algorithm 5. For any $j \in [n], h' < t$, if it holds that $\sigma(\widetilde{\tau}_{j,h'}) \subseteq \sigma(\widetilde{c}_t)$ in $\mathcal{D}_{\mathcal{L}}$ and $\widetilde{a}_{j,h'} \notin \widetilde{\tau}_{i,t}$, then it can compute the associated $\breve{\tau}_{j,h'}$ from induction hypothesis (Line 5-6), compute the exact $\breve{a}_{j,h'}$ based on $\breve{g}_{j,h'}$ (Line 7), and add it into $\breve{\tau}'_{i,h}$ (Line 8). Therefore, we know $\breve{\tau}'_{i,t} = \widetilde{\tau}_{i,t} \cup \{\widetilde{a}_{j,h'} \,|\, \sigma(\widetilde{\tau}_{j,h'}) \subseteq \sigma(\widetilde{c}_t) \text{ and } \widetilde{a}_{j,h'} \notin \widetilde{\tau}_{i,t}\} = \breve{\tau}_{i,t}$.
From induction, we complete the proof. □

Remark .9. The difference between Algorithm 3 and 4 lies as follows. Given any $\breve{g}_{1:\breve{H}}$ and $\mathcal{D}_{\mathcal{L}}$, Algorithm 3 needs to recover the output $\widetilde{a}_{i,h}$ of $\widetilde{g}_{i,h}$ under all possible input $\widetilde{\tau}_{i,h} \in \widetilde{\mathcal{T}}_{i,h}, i \in [n], h \in [\widetilde{H}]$, where the cardinality of $\widetilde{\mathcal{T}}_{i,h}$ could be exponentially large. Thus Algorithm 3 may suffer from computational intractability. However, Algorithm 4 only requires to recover the output $\widetilde{g}_{i,h}$ under the specific $\widetilde{\tau}_{i,h}$ happening in the trajectory, which can be implemented in polynomial time.

4) Proof of Theorem IV.5:

Proof. To prove that $\mathcal{D}^{\dagger}_{\mathcal{L}}$ has SI-CIBs, it suffices to prove that for any $h = 2, \cdots, \breve{H}$, fix any $h_1 \in [h-1], i_1 \in [n]$, and for any $\breve{g}_{1:h-1} \in \breve{\mathcal{G}}_{1:h-1}, \breve{g}'_{i_1,h_1} \in \breve{\mathcal{G}}_{i_1,h_1}$, let $\breve{g}'_{h_1} := (\breve{g}_{1,h_1}, \cdots, \breve{g}'_{i_1,h_1}, \cdots, \breve{g}_{n,h_1})$ and $\breve{g}'_{1:h-1} := (\breve{g}_1, \cdots, \breve{g}'_{h_1}, \cdots, \breve{g}_{h-1})$. If $\breve{c}_h$ is reachable from $\breve{g}_{1:h-1}$ and $\breve{g}'_{1:h-1}$, then the following holds

$$\mathbb{P}(\breve{s}_h, \breve{p}_h \,|\, \breve{c}_h, \breve{g}_{1:h-1}) = \mathbb{P}(\breve{s}_h, \breve{p}_h \,|\, \breve{c}_h, \breve{g}'_{1:h-1}). \tag{.2}$$

We prove this case-by-case as follows:

1) If there exists some $i_3 \neq i_1$ such that $\sigma(\breve{\tau}_{i_1,h_1}) \subseteq \sigma(\breve{\tau}_{i_3,h})$ and $\sigma(\breve{a}_{i_1,h_1}) \subseteq \sigma(\breve{\tau}_{i_3,h})$, then from Assumption II.2, we know that $\sigma(\breve{\tau}_{i_1,h_1}) \subseteq \sigma(\breve{c}_h), \sigma(\breve{a}_{i_1,h_1}) \subseteq \sigma(\breve{c}_h)$. Therefore, there exist deterministic measurable functions α_1, α_2 such that $\breve{\tau}_{i_1,h_1} = \alpha_1(\breve{c}_h), \breve{a}_{i_1,h_1} = \alpha_2(\breve{c}_h)$, and further it holds that

$$\begin{aligned}\mathbb{P}(\breve{s}_h, \breve{p}_h \,|\, \breve{c}_h, \breve{g}_{1:h-1}) &= \mathbb{P}(\breve{s}_h, \breve{p}_h \,|\, \alpha_1(\breve{c}_h), \alpha_2(\breve{c}_h), \breve{c}_h, \breve{g}_{1:h-1}) \\ &= \mathbb{P}(\breve{s}_h, \breve{p}_h \,|\, \breve{\tau}_{i_1,h_1}, \breve{a}_{i_1,h_1}, \breve{c}_h, \breve{g}_{1:h-1}) = \mathbb{P}(\breve{s}_h, \breve{p}_h \,|\, \breve{\tau}_{i_1,h_1}, \breve{a}_{i_1,h_1}, \breve{c}_h, \breve{g}'_{1:h-1}).\end{aligned}$$

The last equality is due to the fact that both the input and output of $\breve{g}_{i_1,h_1}$ are conditioned on.

2) If for any $i_2 \neq i_1$, either $\sigma(\breve{\tau}_{i_1,h_1}) \nsubseteq \sigma(\breve{\tau}_{i_2,h})$ or $\sigma(\breve{a}_{i_1,h_1}) \nsubseteq \sigma(\breve{\tau}_{i_2,h})$, then agent (i_1, h_1) does not influence any agent (i_2, h) with $i_2 \neq i_1$ in $\mathcal{D}^{\dagger}_{\mathcal{L}}$, since otherwise, due to the sQC IS of $\mathcal{D}^{\dagger}_{\mathcal{L}}$, it must hold that $\sigma(\breve{\tau}_{i_1,h_1}) \subseteq \sigma(\breve{\tau}_{i_2,h})$ and $\sigma(\breve{a}_{i_1,h_1}) \subseteq \sigma(\breve{\tau}_{i_2,h})$. Moreover, we claim that such an h_1 has to be even, since otherwise, the agent will be at a communication step, and we must have $\breve{\tau}_{i_1,h_1} = \breve{c}_{h_1} \subseteq \breve{c}_h \subseteq \breve{\tau}_{i_2,h}$ by Assumption III.4 and the reformulation in Eq. (IV.1), and $\breve{a}_{i_1,h_1} = m_{i_1,\frac{h_1+1}{2}} \in z^a_{\frac{h_1+1}{2}} = \breve{z}_{h_1+1} \subseteq \breve{c}_h \subseteq \breve{\tau}_{i_2,h}$ by Assumption II.1 (b), which violates the premise of this case. Let $k_1 := h_1/2$. Now, we claim that agent (i_1, h_1) does not influence the state $\breve{s}_h$ nor the information $\breve{\tau}_{i_1,h}$. We prove this case-by-case as follows:

   a) Suppose h is odd, then $\breve{p}_h = \emptyset$ by Eq. (IV.1). If agent (i_1, h_1) influences $\breve{s}_h$ in $\mathcal{D}^{\dagger}_{\mathcal{L}}$, then agent (i_1, h_1) influences $\widetilde{s}_h$ in $\mathcal{D}_{\mathcal{L}}$ (because strict expansion does not change system dynamics). From Assumption III.7, we know that she also influences $\widetilde{o}_{-i_1,h}$, i.e., there must exist some $i_3 \neq i_1$ such that agent (i_1, h_1) influences $\widetilde{o}_{i_3,h}$ in $\mathcal{D}_{\mathcal{L}}$. From Assumption II.1 (e), it holds that $\widetilde{o}_{i_3,h} \in \widetilde{\tau}_{i_3,h+1}$. Therefore, agent (i_1, h_1) influences agent $(i_3, h+1)$ in $\mathcal{D}_{\mathcal{L}}$. From Lemma .7, we know $\sigma(\tau^-_{i_1,k_1}) \subseteq \sigma(\tau_{i_3,k^-})$ in $\mathcal{L}$, where $k := (h+1)/2$. Furthermore, from Assumption II.2 and $i_3 \neq i_1$, it holds that $\sigma(\tau^-_{i_1,k_1}) \subseteq \sigma(c_{k^-})$. Also, from Eq. (IV.1), it holds that $\widetilde{\tau}_{i_1,h_1} = \tau^+_{i_1,k_1} = \tau^-_{i_1,k_1} \cup z^a_{k_1}$ and $z^a_{k_1} = \widetilde{z}_{h_1} \subseteq \widetilde{c}_h$. Then, we have $\sigma(\widetilde{\tau}_{i_1,h_1}) \subseteq \sigma(\widetilde{c}_h) = \sigma(\widetilde{\tau}_{i_3,h})$. Based on the strict expansion from $\mathcal{D}_{\mathcal{L}}$ to $\mathcal{D}^{\dagger}_{\mathcal{L}}$, we can get $\breve{\tau}_{i_1,h_1} \backslash \widetilde{\tau}_{i_1,h_1} \subseteq \breve{c}_{h_1} \subseteq \breve{\tau}_{i_3,h}$, and $\breve{a}_{i_1,h_1} \in \breve{c}_h$. Then, it holds that $\sigma(\breve{\tau}_{i_1,h_1}) \subseteq \sigma(\breve{\tau}_{i_3,h}), \sigma(\breve{a}_{i_1,h_1}) \subseteq \sigma(\breve{\tau}_{i_3,h})$, which leads to a contradiction to the premise that for any $i_2 \neq i_1$, either $\sigma(\breve{\tau}_{i_1,h_1}) \nsubseteq \sigma(\breve{\tau}_{i_2,h})$ or $\sigma(\breve{a}_{i_1,h_1}) \nsubseteq \sigma(\breve{\tau}_{i_2,h})$. Hence, we know agent (i_1, h_1) does not influence the state $\breve{s}_h$. Additionally, for any $i_2 \neq i_1$, since agent (i_1, h_1) does not influence agent (i_2, h), and $\breve{\tau}_{i_1,h} = \breve{c}_h = \breve{\tau}_{i_2,h}$, then we know that agent (i_1, h_1) does not influence $\breve{\tau}_{i_1,h}$.

   b) Suppose h is even. If agent (i_1, h_1) influences $\breve{s}_{h_1+1}$, then from Assumption III.7, agent (i_1, h_1) influences $\breve{o}_{-i_1,h_1+1}$, and then there exists $i_3 \neq i_1$ such that agent (i_1, h_1) influence $\breve{o}_{i_3,h_1+1}$. However, from Assumption II.1 (e), we know that $\breve{o}_{i_3,h_1+1} = o_{i_3,\frac{h_1}{2}+1} \in \tau_{i_3,\frac{h_1}{2}+1} \subseteq \tau_{i_3,\frac{h}{2}} = \widetilde{\tau}_{i_3,h} \subseteq \breve{\tau}_{i_3,h}$, which means agent (i_1, h_1) influences agent (i_3, h) and leads to a contradiction. Therefore, we know that agent (i_1, h_1) does not influence $\breve{s}_{h_1+1}$, and

further does not influence $\breve{s}_h$. Now, we want to prove that agent (i_1,h_1) does not influence $\breve{\tau}_{i_1,h}$. Firstly, from Assumption II.1 (e), reformulation and strict expansion, we know that $\forall h' \le h, i \in [n], \breve{\tau}_{i,h'} \subseteq \breve{\tau}_{i,h}$. Since agent (i_1,h_1) does not influence $\breve{\tau}_{i_2,h}$ for any $i_2 \neq i_1$, agent (i_1,h_1) does not influence $\breve{\tau}_{i_2,h'}$ with $h' < h$, and thus does not influence $\breve{c}_{h'}$. Secondly, from Assumption III.5, $\breve{a}_{i_1,h_1} = a_{i_1,k_1} \notin p_{i_1,t^-}$ and $\breve{a}_{i_1,h_1} \notin p_{i_1,t^+}$ for any $t > k_1$. Then from Assumption II.1, we know $\widetilde{p}_{i_1,h_1+2} = p_{i_1,(k_1+1)^+} = p_{i_1,(k_1+1)^-} \backslash z^a_{i_1,k_1+1} = \xi_{i_1,k_1+1}(p_{i_1,k_1^+}, a_{i_1,k_1}, o_{i_1,k_1}) \backslash z^a_{i_1,k_1+1} = \xi_{i_1,k_1+1}(\widetilde{p}_{i_1,h_1}, \widetilde{a}_{i_1,h_1}, \widetilde{o}_{i_1,h_1}) \backslash z^a_{i_1,k_1+1}$. Also, we know that $z^a_{i_1,k_1+1} \subseteq c_{(k_1+1)^+} \subseteq \breve{c}_{h_1+2}$ is not influenced by agent (i_1,h_1). And ξ_{i_1,k_1+1} is a fixed transformation, and $a_{i_1,k_1} \notin p_{i_1,(k_1+1)^-}$, we can write $\xi_{i_1,k_1+1}(\widetilde{p}_{i_1,h_1}, \widetilde{a}_{i_1,h_1}, \widetilde{o}_{i_1,h_1}) = \widetilde{\xi}_{i_1,k_1+1}(\widetilde{p}_{i_1,h_1}, \widetilde{o}_{i_1,h_1})$. Also $\widetilde{o}_{i_1,h_1}$ generates from $\widetilde{s}_{h_1} = \breve{s}_{h_1}$ is not influenced by agent (i_1,h_1). Therefore, we know that agent (i_1,h_1) does not influence $\widetilde{p}_{i_1,h_1+2}$. Finally, from the definition, we know that agent (i_1,h_2) does not influence $\breve{\tau}_{i_1,h_1+1} = \breve{c}_{h_1+1}$ and thus does not influence $\breve{a}_{i_1,h_1+1}$. Furthermore, from definition, we know that $\breve{p}_{i_1,h_1+2} = \widetilde{p}_{i_1,h_1+2} \backslash \{\widetilde{a}_{i,t} \,|\, t < h_1+2, \sigma(\widetilde{\tau}_{i,t}) \subseteq \sigma(\widetilde{c}_{h_1+2})\}$. The set $\{(i,t) \,|\, t < h_1+2, \sigma(\widetilde{\tau}_{i,t}) \subseteq \sigma(\widetilde{c}_{h_1+2})\}$ is not influenced by agent (i_1,h_1), and $\widetilde{a}_{i_1,t} = \breve{a}_{i_1,t}$ is also not influenced by agent (i_1,h_1) for any $t < h_1+2$. Therefore, we know that agent (i_1,h_1) does not influence $\{\widetilde{a}_{i,t} \,|\, t < h_1+2, \sigma(\widetilde{\tau}_{i,t}) \subseteq \sigma(\widetilde{c}_{h_1+2})\}$, and thus she does not influence $\breve{p}_{i_1,h_1+2}$. Also, agent (i_1,h_2) does not influence $\breve{c}_{h_1+2}$, agent (i_1,h_1) does not influence $\breve{\tau}_{i_1,h_1+2}$ and $\breve{a}_{i_1,h_1+2}$. In this way, we know that agent (i_1,h_1) does not influence $\breve{\tau}_{i_1,h'}$ for any $h' > h_1$.

Therefore, we know agent (i_1,h_1) does not influence $\breve{s}_h$, and does not influence $\breve{\tau}_{i,h}, \forall i \in [n]$, yielding

$$\begin{aligned}&\mathbb{P}(\breve{s}_h, \breve{p}_h \,|\, \breve{c}_h, \breve{g}_{1:h-1}) = \mathbb{P}(\breve{s}_h, \breve{p}_h, \breve{c}_h \,|\, \breve{c}_h, \breve{g}_{1:h-1}) = \mathbb{P}(\breve{s}_h, \breve{\tau}_h \,|\, \breve{c}_h, \breve{g}_{1:h-1})\\ &= \mathbb{P}(\breve{s}_h, \{\breve{\tau}_{i,h}\}_{i\in[n]} \,|\, \breve{c}_h, \breve{g}_{1:h-1}) = \mathbb{P}(\breve{s}_h, \{\breve{\tau}_{i,h}\}_{i\in[n]} \,|\, \breve{c}_h, \breve{g}'_{1:h-1}) = \mathbb{P}(\breve{s}_h, \breve{p}_h \,|\, \breve{c}_h, \breve{g}'_{1:h-1}).\end{aligned}$$

This completes the proof. □

5) Proof of Theorem IV.6:

Proof. Firstly, from the construction of $\mathcal{D}'_{\mathcal{L}}$ and strategy space $\overline{\mathcal{G}}_{1:\overline{H}}$, we know that for any $h \in [H], i \in [n], \overline{\mathcal{C}}_{2h-1} = \breve{\mathcal{C}}_{2h-1}, \overline{\mathcal{A}}_{i,2h-1} = \breve{\mathcal{A}}_{i,2h-1}, \overline{\mathcal{T}}_{i,2h} = \breve{\mathcal{T}}_{i,2h}, \overline{\mathcal{A}}_{i,2h} = \breve{\mathcal{A}}_{i,2h}$. Therefore, $\overline{\mathcal{G}}_{1:\overline{H}} = \breve{\mathcal{G}}_{1:\breve{H}}$, and finding a team optimal strategy of $\mathcal{D}'_{\mathcal{L}}$ in the strategy space $\overline{\mathcal{G}}_{1:\overline{H}}$ is equivalent to finding a team-optimum of $\mathcal{D}^{\dagger}_{\mathcal{L}}$ in the strategy space $\breve{\mathcal{G}}_{1:\breve{H}}$.
Secondly, we will prove that the Dec-POMDP $\mathcal{D}'_{\mathcal{L}}$ satisfies the information evolution rules in the theorem. For each $t \in [H]$, we define the random variables $\widehat{p}_{i,2t-1} = p_{i,t^-}, \widehat{p}_{2t-1} = p_{t^-}$. Recall that in Eq. (IV.1), $\widetilde{p}_{i,2t-1} = \emptyset$ rather than p_{i,t^-}. Then, from Assumption II.1, it holds that, for any $i \in [n], h \in [\overline{H}]$, if $h = 2t-1$ with $t \in [H]$

$$\widetilde{z}_h = \chi_t(\widetilde{p}_{h-1}, \widetilde{a}_{h-1}, \widetilde{o}_h), \qquad \widehat{p}_{i,h} = \xi_{i,t}(\widetilde{p}_{i,h-1}, \widetilde{a}_{i,h-1}, \widetilde{o}_{i,h});$$

if $h = 2t$ with $t \in [H]$, then

$$\widetilde{z}_h = \phi_t(\widehat{p}_{h-1}, \widetilde{a}_{h-1}), \qquad \widetilde{p}_{i,h} = \widehat{p}_{i,h-1} \backslash \phi_{i,t}(\widehat{p}_{i,h-1}, \widetilde{a}_{i,h-1}),$$

where $\chi_t, \xi_{i,t}$ are fixed transformations and $\phi_h, \phi_{i,h}$ are additional-sharing functions. Then, we can construct the $\{\overline{\chi}_{h+1}\}_{h\in[\overline{H}]}, \{\overline{\xi}_{i,h+1}\}_{i\in[n],h\in[\overline{H}]}$ accordingly as follows:

- If $h = 2t-1$ with $t \in [H]$, for any $\overline{p}_{h-1}, \overline{a}_{h-1}, \overline{o}_h$, since $\overline{p}_{h-1} = \breve{p}_{h-1}$ from construction of $\mathcal{D}'_{\mathcal{L}}$, we can select a $\widetilde{p}_{h-1}$ that $\breve{p}_{h-1}$ can be generated from $\widetilde{p}_{h-1}$ through expansion (such $\widetilde{p}_{h-1}$ might not be unique). Then, define $\overline{\chi}_h(\overline{p}_{h-1}, \overline{a}_{h-1}, \overline{o}_h) = \chi_t(\widetilde{p}_{h-1}, \overline{a}_{h-1}, \overline{o}_h) \cup \{\overline{a}_{j,h_1} \,|\, j \in [n], h_1 < h, \overline{a}_{j,h_1} \in \overline{p}_{h-1}, \sigma(\widetilde{\tau}_{j,h_1}) \subseteq \sigma(\widetilde{c}_h)\} \backslash (\widetilde{p}_{h-1} \backslash \overline{p}_{h-1})$. Since χ_t is a fixed transformation and we remove the $\widetilde{p}_{h-1} \backslash \overline{p}_{h-1}$ part from $\overline{z}_h$, the value $\overline{\chi}_h(\overline{p}_{h-1}, \overline{a}_{h-1}, \overline{o}_h)$ is the same no matter what $\widetilde{p}_{h-1}$ we select, and thus such $\overline{\chi}_h$ is well-defined. Similarly, we can define $\overline{\xi}_{i,h}(\overline{p}_{i,h-1}, \overline{a}_{i,h-1}, \overline{o}_{i,h-1}) = \xi_{i,t}(\widetilde{p}_{i,h-1}, \overline{a}_{i,h-1}, \overline{o}_{i,h}) \backslash \{\overline{a}_{i,h_1} \,|\, h_1 < h, \overline{a}_{i,h_1} \in \overline{p}_{i,h-1}, \sigma(\widetilde{\tau}_{i,h_1}) \subseteq \sigma(\widetilde{c}_h)\} \backslash (\widetilde{p}_{i,h-1} \backslash \overline{p}_{i,h-1})$.
- If $h = 2t$ with $t \in [H]$, for any $\overline{p}_{h-1}, \overline{a}_{h-1}$, from the construction of $\mathcal{D}'_{\mathcal{L}}$, we can select a $\widehat{p}_{h-1}$ that $\overline{p}_{h-1}$ can be generated from $\widehat{p}_{h-1} = p_{t^-}$ through expansion (such $\widehat{p}_{h-1}$ might not be unique). Also, it holds that $\overline{o}_h = \emptyset$, then define $\overline{\chi}_h(\overline{p}_{h-1}, \overline{a}_{h-1}, \overline{o}_h) = \phi_t(\widehat{p}_{h-1}, \overline{a}_{h-1}) \cup \{\overline{a}_{j,h_1} \,|\, j \in [n], h_1 < h, \overline{a}_{j,h_1} \in \overline{p}_{h-1}, \sigma(\widetilde{\tau}_{j,h_1}) \subseteq \sigma(\widetilde{c}_h)\} \backslash (\widehat{p}_{h-1} \backslash \overline{p}_{h-1})$. Still, since ϕ_t is the addition-sharing function, which part of $\widehat{p}_{h-1}$ to share only depends on $\overline{a}_{h-1}$, and not depends on the value of $\widehat{p}_{h-1}$, and we remove the $\widehat{p}_{h-1} \backslash \overline{p}_{h-1}$ part from $\overline{z}_h$, the value of $\overline{\chi}_h(\overline{p}_{h-1}, \overline{a}_{h-1}, \overline{o}_h)$ is the same no matter what $\widehat{p}_{h-1}$ we select, and thus such $\overline{\chi}_h$ is well-defined. Similarly, we can define $\overline{\xi}_{i,h}(\overline{p}_{i,h-1}, \overline{a}_{i,h-1}, \overline{o}_{i,h-1}) = \overline{p}_{i,h-1} \backslash \{\overline{a}_{i,h_1} \,|\, h_1 < h, \overline{a}_{i,h_1} \in \overline{p}_{i,h-1}, \sigma(\widetilde{\tau}_{i,h_1}) \subseteq \sigma(\widetilde{c}_h)\} \backslash \phi_{i,t}(\widehat{p}_{i,h-1}, \overline{a}_{i,h-1})$.

Therefore, the common and private information of $\mathcal{D}'_{\mathcal{L}}$ satisfies that

$$\begin{aligned}\overline{c}_h = \overline{c}_{h-1} \cup \overline{z}_h, \quad \overline{z}_h = \overline{\chi}_h(\overline{p}_{h-1}, \overline{a}_{h-1}, \overline{o}_h)\\ \text{for each } i \in [n], \quad \overline{p}_{i,h} = \overline{\xi}_{i,h}(\overline{p}_{i,h-1}, \overline{a}_{i,h-1}, \overline{o}_{i,h}),\end{aligned}$$

with some functions $\{\overline{\chi}_h\}_{h\in[\overline{H}]}, \{\overline{\xi}_{i,h}\}_{i\in[n],h\in[\overline{H}]}$.
Thirdly, we prove that such a Dec-POMDP $\mathcal{D}'_{\mathcal{L}}$ is SI with respect to the strategy space $\overline{\mathcal{G}}_{1:\overline{H}}$. This is equivalent to that for any

$h \in [2:\overline{H}], \overline{s}_h \in \overline{\mathcal{S}}, \overline{p}_h \in \overline{\mathcal{P}}_h, \overline{c}_h \in \overline{\mathcal{C}}_h, i_1 \in [n], h_1 < h, \overline{g}_{1:h-1}, \overline{g}'_{i_1,h_1} \in \overline{\mathcal{G}}_{i_1:h_1}$, let $\overline{g}'_{h_1} := (\overline{g}_{1,h_1}, \cdots, \overline{g}'_{i_1,h_1}, \cdots, \overline{g}_{n,h_1})$ and $\overline{g}'_{1:h-1} := (\overline{g}_1, \cdots, \overline{g}'_{h_1}, \cdots, \overline{g}_{h-1})$. If $\overline{c}_h$ is reachable from both $\overline{g}_{1:h-1}$ and $\overline{g}'_{1:h-1}$, it holds that

$$\mathbb{P}_h^{\mathcal{D}'_{\mathcal{L}}}(\overline{s}_h, \overline{p}_h \,|\, \overline{c}_h, \overline{g}_{1:h-1}) = \mathbb{P}_h^{\mathcal{D}'_{\mathcal{L}}}(\overline{s}_h, \overline{p}_h \,|\, \overline{c}_h, \overline{g}'_{1:h-1}). \tag{.3}$$

We prove this case-by-case. If $h = 2t$ with $t \in [H]$, then from the result of Theorem IV.5, it holds that

$$\mathbb{P}_h^{\mathcal{D}'_{\mathcal{L}}}(\overline{s}_h, \overline{p}_h \,|\, \overline{c}_h, \overline{g}_{1:h-1}) = \mathbb{P}_h^{\mathcal{D}^\dagger_{\mathcal{L}}}(\overline{s}_h, \overline{p}_h \,|\, \overline{c}_h, \overline{g}_{1:h-1}) = \mathbb{P}_h^{\mathcal{D}^\dagger_{\mathcal{L}}}(\overline{s}_h, \overline{p}_h \,|\, \overline{c}_h, \overline{g}'_{1:h-1}) = \mathbb{P}_h^{\mathcal{D}'_{\mathcal{L}}}(\overline{s}_h, \overline{p}_h \,|\, \overline{c}_h, \overline{g}'_{1:h-1}).$$

Therefore, now we consider the case that $h = 2t-1$ with $t \in [H]$.
Suppose h_1 is odd, which means that $\overline{a}_{h_1}$ corresponds to the communication action in previous $\mathcal{L}$. Then it holds that $\overline{c}_{h_1} \subseteq \overline{c}_h, \overline{a}_{i_1,h_1} = m_{i_1,\frac{h_1+1}{2}} \in \overline{c}_h$, then

$$\begin{aligned}
&\mathbb{P}_h^{\mathcal{D}'_{\mathcal{L}}}(\overline{s}_h, \overline{p}_h \,|\, \overline{c}_h, \overline{g}_{1:h-1}) = \mathbb{P}_h^{\mathcal{D}'_{\mathcal{L}}}(\overline{s}_h, \overline{p}_h \,|\, \overline{c}_{h_1}, \overline{a}_{i_1,h_1}, \overline{c}_h, \overline{g}_{1:h-1}) \\
&= \mathbb{P}_h^{\mathcal{D}'_{\mathcal{L}}}(\overline{s}_h, \overline{p}_h \,|\, \overline{c}_{h_1}, \overline{a}_{i_1,h_1}, \overline{c}_h, \overline{g}_{1:h-1} \backslash \overline{g}_{i_1,h_1}) = \mathbb{P}_h^{\mathcal{D}'_{\mathcal{L}}}(\overline{s}_h, \overline{p}_h \,|\, \overline{c}_h, \overline{g}'_{1:h-1}),
\end{aligned}$$

where the second equality is because the input and output of $\overline{g}_{i_1,h_1}$ are $\overline{c}_{h_1}$ and $\overline{a}_{i_1,h_1}$.
Suppose h_1 is even, which means that h_1 is in the control timestep, let $t_1 = \frac{h_1}{2}$. If $\sigma(\overline{\tau}_{i_1,h_1}) \subseteq \sigma(\overline{c}_h)$ and $\sigma(\overline{a}_{i_1,h_1}) \subseteq \sigma(\overline{c}_h)$, then there exist deterministic measurable functions $\overline{\alpha}_1, \overline{\alpha}_2$ such that $\overline{\tau}_{i_1,h_1} = \overline{\alpha}_1(\overline{c}_h), \overline{a}_{i_1,h_1} = \overline{\alpha}_2(\overline{c}_h)$, and further it holds that

$$\begin{aligned}
&\mathbb{P}(\overline{s}_h, \overline{p}_h \,|\, \overline{c}_h, \overline{g}_{1:h-1}) = \mathbb{P}(\overline{s}_h, \overline{p}_h \,|\, \overline{\alpha}_1(\overline{c}_h), \overline{\alpha}_2(\overline{c}_h), \overline{c}_h, \overline{g}_{1:h-1}) \\
&= \mathbb{P}(\overline{s}_h, \overline{p}_h \,|\, \overline{\tau}_{i_1,h_1}, \overline{a}_{i_1,h_1}, \overline{c}_h, \overline{g}_{1:h-1}) = \mathbb{P}(\overline{s}_h, \overline{p}_h \,|\, \overline{\tau}_{i_1,h_1}, \overline{a}_{i_1,h_1}, \overline{c}_h, \overline{g}'_{1:h-1}).
\end{aligned}$$

If $\sigma(\overline{\tau}_{i_1,h_1}) \nsubseteq \sigma(\overline{c}_h)$ or $\sigma(\overline{a}_{i_1,h_1}) \nsubseteq \sigma(\overline{c}_h)$. Since h_1 is even, $\overline{\tau}_{i_1,h_1} = \breve{\tau}_{i_1,h_1}, \overline{a}_{i_1,h_1} = \breve{a}_{i_1,h_1}$. Also, we know $\overline{c}_h = \breve{c}_h$, then it holds $\sigma(\breve{\tau}_{i_1,h_1}) \nsubseteq \sigma(\breve{c}_h)$ or $\sigma(\breve{a}_{i_1,h_1}) \nsubseteq \sigma(\breve{c}_h)$. Firstly, from the sQC of $\mathcal{D}^\dagger_{\mathcal{L}}$, we know that agent (i_1, h_1) does not influence agent (i_2, h) in $\mathcal{D}^\dagger_{\mathcal{L}}$ for any $i_2 \neq i_1$. Then as shown in Theorem IV.5, we know that agent (i_1, h_1) does not influence $\breve{s}_{h_1+1} = \overline{s}_{h_1+1}$ and does not influence $\breve{c}_{h'} = \overline{c}_{h'}$ for any $h' \leq h$. Secondly, from Assumption II.1, we know that for any $i \in [n], p_{i,(t_1+1)^-} = \xi_{i,t_1+1}(p_{i,t_1^+}, a_{i,t_1}, o_{i,t_1+1})$, where ξ_{i,t_1+1} is an fixed transformation. Also, from Assumption III.5, we know thay $a_{i,t_1} \notin p_{i,(t_1+1)^-}$. Therefore, we can write $p_{i,(t_1+1)^-} = \widetilde{\xi}_{i,t_1+1}(p_{i,t_1^+}, o_{i,(t_1+1)})$. From the definition of refinement, we know that $\overline{p}_{i,h_1+1} = \widetilde{\xi}_{i,t_1}(p_{i,t_1^+}, \overline{o}_{i,h_1+1}) \backslash \overline{c}_{h_1+1}$. Since, agent (i_1, h_1) does not influence $\overline{s}_{h_1+1}$ and thus does not influence $\overline{o}_{i,h_1+1}$. Also, agent (i_1, h_1) does not influence $\overline{c}_{h_1+1}$ and p_{i,h_1^+} (happens before choosing $a_{i_1,t_1} = \overline{a}_{i_1,h_1}$). Therefore, agent (i_1, h_1) does not influence $\overline{\tau}_{i,h_1+1}$ and thus does not influence $\overline{a}_{i,h_1+1}$ for any $i \in [n]$. Thirdly, we know that for any $i \in [n], \overline{p}_{i,h_1+2} = \overline{\xi}_{i,h_1+2}(\overline{p}_{i,h_1+1}, \overline{a}_{i,h_1+1}, \overline{o}_{i,h_1+2})$, where $\overline{o}_{i,h_1+2} = \emptyset$. Since agent (i_1, h_1) does not influence $\overline{p}_{i,h_1+1}$ and $\overline{a}_{i,h_1+1}$, then it does not influences $\overline{p}_{i,h_1+2}$. Also we know that it does not influence $\overline{\tau}_{i,h_1+2}$. Therefore, agent (i_1, h_1) does not influence $\overline{\tau}_{i,h_1+2}$ and thus does not influence $\overline{a}_{i,h_1+2}$. In this way, we can have agent (i_1, h_1) does not influence $\overline{\tau}_{i,h'}$ for any $i \in [n], h_1 < h' \leq h$.
Finally, we know agent (i_1, h_1) does not influence $\overline{s}_h$, and does not influence $\overline{\tau}_{i,h}, \forall i \in [n]$, yielding

$$\begin{aligned}
&\mathbb{P}(\overline{s}_h, \overline{p}_h \,|\, \overline{c}_h, \overline{g}_{1:h-1}) = \mathbb{P}(\overline{s}_h, \overline{p}_h, \overline{c}_h \,|\, \overline{c}_h, \overline{g}_{1:h-1}) = \mathbb{P}(\overline{s}_h, \overline{\tau}_h \,|\, \overline{c}_h, \overline{g}_{1:h-1}) \\
&= \mathbb{P}(\overline{s}_h, \{\overline{\tau}_{i,h}\}_{i\in[n]} \,|\, \overline{c}_h, \overline{g}_{1:h-1}) = \mathbb{P}(\overline{s}_h, \{\overline{\tau}_{i,h}\}_{i\in[n]} \,|\, \overline{c}_h, \overline{g}'_{1:h-1}) = \mathbb{P}(\overline{s}_h, \overline{p}_h \,|\, \overline{c}_h, \overline{g}'_{1:h-1}).
\end{aligned}$$

which completes the proof. □

6) Important Definitions of SI-CIB Dec-POMDP: Given an SI-CIB Dec-POMDP $\mathcal{D}'_{\mathcal{L}}$ obtained from $\mathcal{L}$ after reformulation, strict expansion and refinement. In this part, we only need to discuss how to solve this $\mathcal{D}'_{\mathcal{L}}$. Recall that we use $\bar{}$ for the notation of the elements and quantities in $\mathcal{D}'_{\mathcal{L}}$.

First, we define the following quantities, and most of them are adapted from [14].

Definition .10 (Value function)**.** For each $i \in [n]$ and $h \in [\overline{H}]$, given common information $\overline{c}_h$ and strategy $\overline{g}_{1:H}$, the value function conditioned on the common information is defined as:

$$V_h^{\overline{g},\mathcal{D}'_{\mathcal{L}}}(\overline{c}_h) := \mathbb{E}_{\overline{g}}^{\mathcal{D}'_{\mathcal{L}}}\left[\sum_{h'=h}^{\overline{H}} \overline{\mathcal{R}}_{h'}(\overline{s}_{h'}, \overline{a}_{h'}, \overline{p}_{h'}) \,|\, \overline{c}_h\right], \tag{.4}$$

where $\overline{\mathcal{R}}_{h'}$ takes $\overline{s}_{h'}, \overline{a}_{h'}, \overline{p}_{h'}$ as input, since after reformulation, the reward may come from communication cost, which is a function of $\overline{p}_{h'}$ and $\overline{a}_{h'}$.

Definition .11 (Prescription and Q-Value function)**.** Prescription is an important concept in the common-information-based framework [15], [16]. The prescription of agent i at the timestep h is defined as $\gamma_{i,h} : \overline{\mathcal{P}}_{i,h} \to \overline{\mathcal{A}}_{i,h}$. We use $\gamma_h := (\gamma_{1,h}, \cdots, \gamma_{n,h})$ to denote the joint prescription and $\Gamma_{i,h}, \Gamma_h$ to denote the prescription space. The prescriptions are the marginalization of strategy $\overline{g}_h$, i.e., $\gamma_{i,h}(\cdot \,|\, \overline{p}_{i,h}) = \overline{g}_{i,h}(\cdot \,|\, \overline{c}_h, \overline{p}_{i,h})$. Then we can define the Q-value function as

$$Q_h^{\overline{g}, \mathcal{D}'_{\mathcal{L}}}(\overline{c}_h, \gamma_h) := \mathbb{E}_{\overline{g}}^{\mathcal{D}'_{\mathcal{L}}} \left[\sum_{h'=h}^{\overline{H}} \overline{\mathcal{R}}_{h'}(\overline{s}'_h, \overline{a}'_h, \overline{p}'_h) \,|\, \overline{c}_h, \gamma_h \right]. \tag{.5}$$

Remark .12. In this paper, for any Dec-POMDP $\mathcal{D}'_{\mathcal{L}}$ generated by an $\mathcal{L}$ after reformulation, strict expansion, and refinement, we only consider the strategy spaces at odd timesteps as $\overline{\mathcal{G}}_{i,2t-1} : \overline{\mathcal{C}}_{2t-1} \to \overline{\mathcal{A}}_{i,2t-1}$ and aim to find the optimal strategy in these classes. Therefore, we define the prescription spaces at odd timesteps as $\forall h \in [H], i \in [n], \Gamma_{i,2h-1} = \overline{\mathcal{A}}_{i,2h-1} = \mathcal{M}_{i,h}, \Gamma_{2h-1} = \overline{\mathcal{A}}_{2h-1} = \mathcal{M}_h$.

Definition .13 (Expected approximate common information model)**.** We define an expected approximate common information model of $\mathcal{D}'_{\mathcal{L}}$ as

$$\mathcal{M} := \left(\{\widehat{\mathcal{C}}_h\}_{h \in [\overline{H}]}, \{\widehat{\phi}_h\}_{h \in [\overline{H}]}, \{\mathbb{P}_h^{\mathcal{M}, z}\}_{h \in [\overline{H}]}, \Gamma, \{\widehat{\mathcal{R}}_h^{\mathcal{M}}\}_{h \in [\overline{H}]} \right), \tag{.6}$$

where Γ is the joint prescription space, $\widehat{\mathcal{C}}_h$ is the space of approximate common information at step h. $\mathbb{P}_h^{\mathcal{M}, z} : \widehat{\mathcal{C}}_h \times \Gamma_h \to \Delta(\overline{Z}_{h+1})$ gives the probability of $\overline{z}_{h+1}$ under $\widehat{c}_h$ and γ_h. $\widehat{\mathcal{R}}_h^{\mathcal{M}} : \widehat{\mathcal{C}}_h \times \Gamma_h \to [0, 1]$ gives the reward at timestep h given $\widehat{c}_h$ and γ_h. Then, we call that $\mathcal{M}$ is an $(\epsilon_r(\mathcal{M}), \epsilon_z(\mathcal{M}))$-expected-approximate common information model of $\mathcal{D}'_{\mathcal{L}}$ with some compression function Compress_h such that $\widehat{c}_h = \text{Compress}_h(\overline{c}_h)$ satisfies the following:

- There exists a transformation function $\widehat{\phi}_h$ such that

$$\widehat{c}_h = \widehat{\phi}_h(\widehat{c}_{h-1}, \overline{z}_h), \tag{.7}$$

  where $\overline{z}_h = \overline{c}_h \backslash \overline{c}_{h-1}$ in $\mathcal{D}'_{\mathcal{L}}$.
- For any $\overline{g}_{1:h-1}$ and any prescription $\gamma_h \in \Gamma_h$, it holds that

$$\mathbb{E}_{\overline{a}_{1:h-1}, \overline{o}_{1:h} \sim \overline{g}_{1:h-1}}^{\mathcal{D}'_{\mathcal{L}}} |\mathbb{E}^{\mathcal{D}'_{\mathcal{L}}}[\overline{\mathcal{R}}_h(\overline{s}_h, \overline{a}_h, \overline{p}_h) \,|\, \overline{c}_h, \gamma_h] - \widehat{\mathcal{R}}_h^{\mathcal{M}}(\widehat{c}_h, \gamma_h)| \leq \epsilon_r(\mathcal{M}). \tag{.8}$$

- For any $\overline{g}_{1:h-1}$ and any prescription $\gamma_h \in \Gamma_h$, it holds that

$$\mathbb{E}_{\overline{a}_{1:h-1}, \overline{o}_{1:h} \sim \overline{g}_{1:h-1}}^{\mathcal{D}'_{\mathcal{L}}} ||\mathbb{P}_h^{\mathcal{D}'_{\mathcal{L}}}(\cdot \,|\, \overline{c}_h, \gamma_h) - \mathbb{P}_h^{\mathcal{M}, z}(\cdot \,|\, \widehat{c}_h, \gamma_h)||_1 \leq \epsilon_z(\mathcal{M}). \tag{.9}$$

Definition .14 (Value function under $\mathcal{M}$)**.** Given a Dec-POMDP $\mathcal{D}'_{\mathcal{L}}$ and its expected approximate common information model $\mathcal{M}$. For any strategy $\overline{g}_{1:\overline{H}} \in \overline{\mathcal{G}}_{1:\overline{H}}, h \in [\overline{H}]$, we define the value function as

$$V_h^{\overline{g}_{1:\overline{H}}, \mathcal{M}}(\overline{c}_h) = \widehat{\mathcal{R}}_h^{\mathcal{M}}(\text{Compress}_h(\overline{c}_h), \{\overline{g}_{j,h}(\cdot \,|\, \overline{c}_h, \cdot)\}_{j \in [n]}) + \mathbb{E}^{\mathcal{M}}[V_h^{\overline{g}_{1:\overline{H}}, \mathcal{M}}(\overline{c}_{h+1}) \,|\, \text{Compress}_h(\overline{c}_h), \{\overline{g}_{j,h}(\cdot \,|\, \overline{c}_h, \cdot)\}_{j \in [n]}]. \tag{.10}$$

Definition .15 (Model-belief consistency)**.** We say the expected approximate common information model $\mathcal{M}$ is *consistent with* some belief $\{\mathbb{P}_h^{\mathcal{M}, c}(\overline{s}_h, \overline{p}_h \,|\, \widehat{c}_h)\}_{h \in [H]}$ if it satisfies the following for all $i \in [n]$, $h \in [H]$:

$$\mathbb{P}_h^{\mathcal{M}, z}(\overline{z}_{h+1} \,|\, \widehat{c}_h, \gamma_h) = \sum_{\substack{\overline{s}_h, \overline{p}_h, \overline{a}_h, \overline{o}_{h+1}: \\ \chi_{h+1}(\overline{p}_h, \overline{a}_h, \overline{o}_{h+1}) = \overline{z}_{h+1}}} \left(\mathbb{P}_h^{\mathcal{M}, c}(\overline{s}_h, \overline{p}_h \,|\, \widehat{c}_h) \mathbb{1}[\overline{a}_h = \gamma_h(\overline{p}_h)] \sum_{s_{h+1}} \overline{\mathbb{T}}_h(\overline{s}_{h+1} \,|\, \overline{s}_h, \overline{a}_h)] \overline{\mathbb{O}}_{h+1}(\overline{o}_{h+1} \,|\, \overline{s}_{h+1}) \right), \tag{.11}$$

$$\widehat{\mathcal{R}}_h^{\mathcal{M}}(\widehat{c}_h, \gamma_h) = \sum_{\overline{s}_h, \overline{p}_h, \overline{a}_h} \mathbb{P}_h^{\mathcal{M}, c}(\overline{s}_h, \overline{p}_h \,|\, \widehat{c}_h) \mathbb{1}[\overline{a}_h = \gamma_h(\overline{p}_h)] \overline{\mathcal{R}}_h(\overline{s}_h, \overline{a}_h, \overline{p}_h). \tag{.12}$$

Definition .16 (Strategy-dependent approximate common information model)**.** Given a model $\widetilde{\mathcal{M}}$ (as in Definition .13) and H joint strategies $g^{1:\overline{H}}$, where each $g^h \in \overline{\mathcal{G}}_{1:\overline{H}}$ for $h \in [\overline{H}]$, we say $\widetilde{\mathcal{M}}$ is a *strategy-dependent expected approximate common information model*, denoted as $\widetilde{\mathcal{M}}(\pi^{1:H})$, if it is consistent with the *strategy-dependent* belief $\{\mathbb{P}_h^{\pi^h, \mathcal{D}'_{\mathcal{L}}}(s_h, p_h \,|\, \widehat{c}_h)\}_{h \in [H]}$ (as per Definition .15). we say $\widetilde{\mathcal{M}}$ is a *strategy-dependent expected approximate common information model*, denoted as $\widetilde{\mathcal{M}}(g^{1:H})$, if it is consistent with the *strategy-dependent* belief $\{\mathbb{P}_h^{g^h, \mathcal{D}'_{\mathcal{L}}}(\overline{s}_h, \overline{p}_h \,|\, \widehat{c}_h)\}_{h \in [\overline{H}]}$ (as per Definition .15).

Definition .17 (Length of approximate common information)**.** Given the compression functions $\{\text{Compress}_h\}_{h \in [\overline{H}+1]}$, we define the integer $\widehat{L} > 0$ as the minimum length such that there exists a mapping $\widehat{f}_h : \overline{\mathcal{A}}_{\max\{1, h-\widehat{L}\}:h-1} \times \overline{\mathcal{O}}_{\max\{1, h-\widehat{L}+1\}, h} \to \widehat{\mathcal{C}}_h$ such that for each $h \in [\overline{H}+1]$ and joint history $\{\overline{o}_{1:h}, \overline{a}_{1:h-1}\}$, we have $\widehat{f}_h(x_h) = \widehat{c}_h$, where $x_h = \{\overline{a}_{\max\{h-\widehat{L}, 1\}}, \overline{o}_{\max\{h-\widehat{L}, 1\}+1}, \cdots, \overline{a}_{h-1}, \overline{o}_h\}$.

7) Main Results for Planning in QC LTC: Finally, we provide the formal guarantees for planning in QC LTC.

Theorem .18. Given any QC LTC problem $\mathcal{L}$ satisfying Assumptions III.1, III.4, III.5, III.7, and IV.7, we can construct an SI Dec-POMDP problem $\mathcal{D}'_{\mathcal{L}}$ such that for any $\epsilon > 0$, solving an ϵ-team optimal strategy in $\mathcal{D}'_{\mathcal{L}}$ can give us an ϵ-team optimal strategy of $\mathcal{L}$, and the following holds. Fix $\epsilon_r, \epsilon_z > 0$ and given any (ϵ_r, ϵ_z)-expected-approximate common information model $\mathcal{M}$ for $\mathcal{D}'_{\mathcal{L}}$ that is consistent with some given approximate belief $\{\mathbb{P}_h^{\mathcal{M},c}(\overline{s}_h, \overline{p}_h \,|\, \widehat{c}_h)\}_{h\in[\overline{H}]}$, Algorithm 1 can compute a $(2\overline{H}\epsilon_r + \overline{H}^2\epsilon_z)$-team optimal strategy for the original LTC problem $\mathcal{L}$ with time complexity $\max_{h\in[\overline{H}]} |\widehat{\mathcal{C}}_h| \cdot \mathtt{poly}(|\mathcal{S}|, |\mathcal{A}_h|, |\mathcal{P}_h|, \overline{H})$. In particular, for fixed $\epsilon > 0$, if $\mathcal{L}$ has any one of baseline sharing protocols as in Appendix A, one can construct a $\mathcal{M}$ and apply Algorithm 1 to compute an ϵ-team optimal strategy for $\mathcal{L}$ in quasi-polynomial time.

Proof. We divide the proof into the following three **Parts**.

Part I: Given any QC LTC problem $\mathcal{L}$ satisfying Assumptions III.1, III.4, III.5, and III.7, we can construct an SI Dec-POMDP problem $\mathcal{D}'_{\mathcal{L}}$ such that finding an ϵ-team optimal strategy can give us an ϵ-team optimal strategy of $\mathcal{L}$, as shown in Algorithm 1.
We can construct a Dec-POMDP $\mathcal{D}'_{\mathcal{L}}$ from $\mathcal{L}$ through Algorithm 1. From Proposition IV.1 and Theorems IV.4, IV.5. We know that $\mathcal{D}'_{\mathcal{L}}$ is SI and an ϵ-team-optimal strategy of $\mathcal{D}'_{\mathcal{L}}$ can give us an ϵ-team optimal strategy of $\mathcal{L}$.

Part II: Given any ϵ-expected-approximate common information model $\mathcal{M}$ of the Dec-POMDP $\mathcal{D}'_{\mathcal{L}}$, there exists an algorithm, Algorithm 6, that can output an ϵ-team optimal strategy of $\mathcal{D}'_{\mathcal{L}}$.
First, we need to prove that solving $\mathcal{M}$ can get the ϵ-team optimal strategy of $\mathcal{D}'_{\mathcal{L}}$. We prove the following 2 lemmas first.

Lemma .19. For any strategy $\overline{g}_{1:\overline{H}} \in \overline{\mathcal{G}}_{1:\overline{H}}$, and $h \in [\overline{H}]$, we have

$$\mathbb{E}^{\mathcal{D}'_{\mathcal{L}}}_{\overline{g}_{1:\overline{H}}}[|V_h^{\overline{g}_{1:\overline{H}},\mathcal{D}'_{\mathcal{L}}}(\overline{c}_h) - V_h^{\overline{g}_{1:\overline{H}},\mathcal{M}}(\overline{c}_h)|] \le (\overline{H}-h+1)\epsilon_r + \frac{(\overline{H}-h+1)(\overline{H}-h)}{2}\epsilon_z. \tag{.13}$$

Proof. We prove it by induction. For $h = \overline{H}+1$, we have $V_h^{\overline{g}_{1:\overline{H}},\mathcal{D}'_{\mathcal{L}}}(\overline{c}_h) = V_h^{\overline{g}_{1:\overline{H}},\mathcal{M}}(\overline{c}_h) = 0$.
For the step $h \le \overline{H}$, we have

$$\begin{aligned}
&\mathbb{E}^{\mathcal{D}'_{\mathcal{L}}}_{\overline{g}_{1:\overline{H}}}[|V_h^{\overline{g}_{1:\overline{H}},\mathcal{D}'_{\mathcal{L}}}(\overline{c}_h) - V_h^{\overline{g}_{1:\overline{H}},\mathcal{M}}(\overline{c}_h)|] \\
\le&\mathbb{E}^{\mathcal{D}'_{\mathcal{L}}}_{\overline{g}_{1:\overline{H}}}\left[|\mathbb{E}^{\mathcal{D}_{\mathcal{L}}}[\overline{\mathcal{R}}_h(\overline{s}_h,\overline{a}_h,\overline{p}_h) \,|\, \overline{c}_h, \{\overline{g}_{j,h}(\cdot \,|\, \overline{c}_h,\cdot)\}_{j\in[n]}] - \widehat{\mathcal{R}}_h^{\mathcal{M}}(\widehat{c}_h, \{\overline{g}_{j,h}(\cdot \,|\, \overline{c}_h,\cdot)\}_{j\in[n]})|\right] \\
&+ \mathbb{E}^{\mathcal{D}'_{\mathcal{L}}}_{\overline{g}_{1:\overline{H}}}\left[|\mathbb{E}_{\overline{z}_{h+1}\sim\mathbb{P}_h^{\mathcal{D}'_{\mathcal{L}}}(\cdot \,|\, \overline{c}_h, \{\overline{g}_{j,h}(\cdot \,|\, \overline{c}_h,\cdot)\}_{j\in[n]})}[V_h^{\overline{g}_{1:\overline{H}},\mathcal{D}'_{\mathcal{L}}}(\overline{c}_h \cup \overline{z}_{h+1})] - \mathbb{E}_{\overline{z}_{h+1}\sim\mathbb{P}_h^{\mathcal{M},z}(\cdot \,|\, \widehat{c}_h, \{\overline{g}_{j,h}(\cdot \,|\, \overline{c}_h,\cdot)\}_{j\in[n]})}[V_h^{\overline{g}_{1:\overline{H}},\mathcal{M}}(\overline{c}_h \cup \overline{z}_{h+1})]|\right] \\
\le&\epsilon_r + (\overline{H}-h)\mathbb{E}^{\mathcal{D}'_{\mathcal{L}}}_{\overline{a}_{1:h-1},\overline{o}_{1:h}\sim\overline{g}_{1:h-1}}||\mathbb{P}_h^{\mathcal{D}'_{\mathcal{L}}}(\cdot \,|\, \overline{c}_h,\gamma_h) - \mathbb{P}_h^{\mathcal{M},z}(\cdot \,|\, \widehat{c}_h,\gamma_h)||_1 + \mathbb{E}^{\mathcal{D}'_{\mathcal{L}}}_{\overline{a}_{1:h-1},\overline{o}_{1:h}\sim\overline{g}_{1:h-1}}\left[|V_{h+1}^{\overline{g}_{1:\overline{H}},\mathcal{D}'_{\mathcal{L}}}(\overline{c}_{h+1}) - V_{h+1}^{\overline{g}_{1:\overline{H}},\mathcal{M}}(\overline{c}_{h+1})|\right] \\
\le&\epsilon_r + (\overline{H}-h)\epsilon_z + (\overline{H}-h)\epsilon_r + \frac{(\overline{H}-h)(\overline{H}-h-1)}{2}\epsilon_z \\
\le&(\overline{H}-h+1)\epsilon_r + \frac{(\overline{H}-h)(\overline{H}-h+1)}{2}\epsilon_z.
\end{aligned}$$

The proof mainly follows from the proof of Lemma 2 in [14]. But the difference is that $\mathcal{D}'_{\mathcal{L}}$ may not satisfy Assumption II.1. In the third line of this proof, we had $\overline{z}_{h+1} \sim \mathbb{P}_h^{\mathcal{D}'_{\mathcal{L}}}(\cdot \,|\, \overline{c}_h, \{\overline{g}_{j,h}(\cdot \,|\, \overline{c}_h,\cdot)\}_{j\in[n]})$, where $\overline{z}_{h+1}$ is generated as

$$\begin{aligned}
&\mathbb{P}_h^{\mathcal{D}'_{\mathcal{L}}}(\overline{z}_{h+1} \,|\, \overline{c}_h,\gamma_h) \\
=&\sum_{\overline{s}_h\in\overline{\mathcal{S}},\overline{p}_h\in\overline{\mathcal{P}}_h} \mathbb{P}_h^{\mathcal{D}'_{\mathcal{L}}}(\overline{s}_h, \quad \overline{p}_h \,|\, \overline{c}_h) \sum_{\overline{s}_{h+1}\in\overline{\mathcal{S}},\overline{o}_{h+1}\in\overline{\mathcal{O}}_{h+1}} \overline{\mathbb{T}}_{h+1}(\overline{s}_{h+1} \,|\, \overline{s}_h, \gamma_h(\overline{p}_h))\overline{\mathbb{O}}_{h+1}(\overline{o}_{h+1} \,|\, \overline{s}_{h+1})\mathbb{1}[\overline{\chi}_{h+1}(\overline{p}_h, \gamma_h(\overline{p}_h), \overline{o}_{h+1})],
\end{aligned}$$

with $\gamma_h = \{\overline{g}_{j,h}(\cdot \,|\, \overline{c}_h,\cdot)\}_{j\in[n]}$. □

Lemma .20. Let $\widehat{g}^*_{1:\overline{H}} \in \overline{\mathcal{G}}_{1:\overline{H}}$ be the strategy output by Algorithm 6, then for any $h \in [\overline{H}], \overline{c}_h \in \overline{\mathcal{C}}_h, \overline{g}_{1:\overline{H}} \in \overline{\mathcal{G}}_{1:\overline{H}}$, it holds that

$$V_h^{\overline{g}_{1:\overline{H}},\mathcal{M}}(\overline{c}_h) \le V_h^{\widehat{g}^*_{1:\overline{H}},\mathcal{M}}(\overline{c}_h). \tag{.14}$$

Proof. We prove it by induction. For $h = \overline{H}+1$, we have $V_h^{\overline{g}_{1:\overline{H}},\mathcal{M}}(\overline{c}_h) = V_h^{\widehat{g}^*_{1:\overline{H}},\mathcal{M}}(\overline{c}_h) = 0$.

For the timestep $h \leq H$, we have

$$\begin{aligned}
V_h^{\overline{g}_{1:\overline{H}},\mathcal{M}}(\overline{c}_h) &= \mathbb{E}^{\mathcal{M}}[\widehat{r}_h^{\mathcal{M}}(\widehat{c}_h, \{\overline{g}_{j,h}(\cdot \,|\, \overline{c}_h, \cdot)\}_{j\in[n]}) + V_{h+1}^{\overline{g}_{1:\overline{H}},\mathcal{M}}(\overline{c}_{h+1}) \,|\, \widehat{c}_h, \overline{g}_h] \\
&\leq \mathbb{E}^{\mathcal{M}}[\widehat{r}_h^{\mathcal{M}}(\widehat{c}_h, \{\overline{g}_{j,h}(\cdot \,|\, \overline{c}_h, \cdot)\}_{j\in[n]}) + V_{h+1}^{\widehat{g}^*_{1:\overline{H}},\mathcal{M}}(\overline{c}_{h+1}) \,|\, \widehat{c}_h, \overline{g}_h] \\
&= Q_h^{\widehat{g}^*_{1:\overline{H}},\mathcal{M}}(\overline{c}_h, \{\overline{g}_{j,h}(\cdot \,|\, \overline{c}_h)\}_{j\in[n]}) \\
&\leq Q_h^{\widehat{g}^*_{1:\overline{H}},\mathcal{M}}(\overline{c}_h, \{\widehat{g}_{j,h}(\cdot \,|\, \overline{c}_h)\}_{j\in[n]}) \\
&= V_h^{\widehat{g}^*_{1:\overline{H}},\mathcal{M}}(\overline{c}_h).
\end{aligned}$$

For the first inequality, we use the induction hypothesis. For the second inequality sign, we use the property of argmax in algorithm and $V_h^{\widehat{g}^*_{1:\overline{H}},\mathcal{M}}(\overline{c}_h) = V_h^{\widehat{g}^*_{1:\overline{H}},\mathcal{M}}(\widehat{c}_h)$. By induction, we complete the proof. □

We now go back to the proof of the theorem. Let $\widehat{g}^*_{1:\overline{H}}$ be the solution output by Algorithm 6, then for any $\overline{g}_{1:\overline{H}} \in \overline{\mathcal{G}}_{1:\overline{H}}, h \in [\overline{H}], \overline{c}_h \in \overline{\mathcal{C}}_h$, we have

$$\begin{aligned}
&\mathbb{E}^{\mathcal{D}'_{\mathcal{L}}}_{\overline{g}_{1:\overline{H}}}\left[V_h^{\overline{g}_{1:\overline{H}},\mathcal{D}'_{\mathcal{L}}}(\overline{c}_h) - V_h^{\widehat{g}^*_{1:\overline{H}},\mathcal{D}'_{\mathcal{L}}}(\overline{c}_h)\right] \\
&= \mathbb{E}^{\mathcal{D}'_{\mathcal{L}}}_{\overline{g}_{1:\overline{H}}}\left[\left(V_h^{\overline{g}_{1:\overline{H}},\mathcal{D}'_{\mathcal{L}}}(\overline{c}_h) - V_h^{\widehat{g}^*_{1:\overline{H}},\mathcal{M}}(\overline{c}_h)\right) + \left(V_h^{\widehat{g}^*_{1:\overline{H}},\mathcal{M}}(\overline{c}_h) - V_h^{\widehat{g}^*_{1:\overline{H}},\mathcal{D}'_{\mathcal{L}}}(\overline{c}_h)\right)\right] \\
&\leq \mathbb{E}^{\mathcal{D}'_{\mathcal{L}}}_{\overline{g}_{1:\overline{H}}}\left[\left(V_h^{\overline{g}_{1:\overline{H}},\mathcal{D}'_{\mathcal{L}}}(\overline{c}_h) - V_h^{\overline{g}_{1:\overline{H}},\mathcal{M}}(\overline{c}_h)\right) + \left(V_h^{\widehat{g}^*_{1:\overline{H}},\mathcal{M}}(\overline{c}_h) - V_h^{\widehat{g}^*_{1:\overline{H}},\mathcal{D}'_{\mathcal{L}}}(\overline{c}_h)\right)\right] \\
&\leq (\overline{H}-h+1)\epsilon_r + \frac{(\overline{H}-h)(\overline{H}-h+1)}{2}\epsilon_z + (\overline{H}-h+1)\epsilon_r + \frac{(\overline{H}-h)(\overline{H}-h+1)}{2}\epsilon_z \\
&= 2(\overline{H}-h+1)\epsilon_r + (\overline{H}-h)(\overline{H}-h+1)\epsilon_z.
\end{aligned} \tag{.15}$$

For the first inequality, we use Lemma .20. For the second inequality sign, we use Lemma .19. Then apply $h = 1$, we have $J_{\mathcal{D}'_{\mathcal{L}}}(\overline{g}_{1:\overline{H}}) \leq J_{\mathcal{D}'_{\mathcal{L}}}(\widehat{g}^*_{1:\overline{H}}) + 2\overline{H}\epsilon_r + \overline{H}^2\epsilon_z$. This completes the proof of **Part II**.

Part III: If the baseline sharing of $\mathcal{L}$ is one of the 4 cases in §A, we can construct an expected-approximate common information model of $\mathcal{D}'_{\mathcal{L}}$.
3kkk We first prove following lemmas: We aim to bound (ϵ_r, ϵ_z) using the following lemma.

Lemma .21. Given any belief $\{\mathbb{P}_h^{\mathcal{M},c}(\overline{s}_h, \overline{p}_h\}_{h\in[\overline{H}]}$ consistent with the expected-approximate-common-information model $\mathcal{M}$, it holds that for any $h \in [\overline{H}], \overline{\mathcal{C}}_h, \gamma_h \in \Gamma_h$:

$$\begin{aligned}
||\mathbb{P}_h^{\mathcal{D}'_{\mathcal{L}}}(\cdot \,|\, \overline{c}_h, \gamma_h) - \mathbb{P}_h^{\mathcal{M},z}(\cdot \,|\, \widehat{c}_h, \gamma_h)||_1 &\leq ||\mathbb{P}_h^{\mathcal{D}'_{\mathcal{L}}}(\cdot, \cdot \,|\, \overline{c}_h) - \mathbb{P}_h^{\mathcal{M},c}(\cdot, \cdot \,|\, \widehat{c}_h)||_1, \\
|\mathbb{E}^{\mathcal{D}'_{\mathcal{L}}}[\mathcal{R}_h(\overline{s}_h, \overline{a}_h, \overline{p}_h) \,|\, \overline{c}_h, \gamma_h] - \widehat{\mathcal{R}}_h^{\mathcal{M}}(\widehat{c}_h, \gamma_h)| &\leq ||\mathbb{P}_h^{\mathcal{D}'_{\mathcal{L}}}(\cdot, \cdot \,|\, \overline{c}_h) - \mathbb{P}_h^{\mathcal{M},c}(\cdot, \cdot \,|\, \widehat{c}_h)||_1,
\end{aligned}$$

where $\widehat{c}_h$ =Compress$_h(\overline{c}_h)$.

Proof. Adapted from Lemma 3 in [14] by changing the reward function of $r_{i,h}(s_h, a_h)$ to $\mathcal{R}_h(\overline{s}_h, \overline{a}_h, \overline{p}_h)$. Note that the latter can still be evaluated given the common-information-based belief, $\mathbb{P}_h^{\mathcal{D}'_{\mathcal{L}}}(\overline{s}_h, \overline{p}_h \,|\, \overline{c}_h)$. □

Then we define the belief states following the notation in [25], [14] as $\overline{\boldsymbol{b}}_1(\emptyset) = \mu_1$, $\overline{\boldsymbol{b}}_h(\overline{a}_{1:h-1}, \overline{o}_{1:h}) = \mathbb{P}(\overline{s}_h = \cdot \,|\, \overline{o}_{1:h}, \overline{a}_{1:h-1}), \overline{\boldsymbol{b}}_h(\overline{a}_{1:h-1}, \overline{o}_{1:h-1}) = \mathbb{P}(\overline{s}_h = \cdot \,|\, \overline{o}_{1:h-1}, \overline{a}_{1:h-1})$, where $\overline{\boldsymbol{b}} \in \Delta(\mathcal{S})$. Also, we define the approximate belief state using the most recent L-step history, that

$$\begin{aligned}
\overline{\boldsymbol{b}}'_h(\overline{a}_{h-L:h-1}, \overline{o}_{h-L+1:h-1}) &= \mathbb{P}(\overline{s}_h = \cdot \,|\, \overline{s}_{h-L} \sim \text{Unif}(\mathcal{S}), \overline{a}_{h-L:h-1}, \overline{o}_{h-L+1:h}) \\
\overline{\boldsymbol{b}}'_h(\overline{a}_{h-L:h-1}, \overline{o}_{h-L+1:h-1}) &= \mathbb{P}(\overline{s}_h = \cdot \,|\, \overline{s}_{h-L} \sim \text{Unif}(\mathcal{S}), \overline{a}_{h-L:h-1}, \overline{o}_{h-L+1:h}).
\end{aligned}$$

Also, for any set $N \subseteq [n]$, we define $\overline{a}_{N,h} = \{\overline{a}_{i,h}\}_{i\in N}$, and the same for $\overline{o}_{N,h}$. We can also define the belief of states given historical observations and actions as follows: for any $N \subseteq [n]$,

$$\begin{aligned}
\overline{\boldsymbol{b}}_h(\overline{a}_{1:h-1}, \overline{o}_{1:h-1}, \overline{o}_{N,h}) &= \mathbb{P}(\overline{s}_h = \cdot \,|\, \overline{a}_{1:h-1}, \overline{o}_{1:h-1}, \overline{o}_{N,h}) \\
\overline{\boldsymbol{b}}'_h(\overline{a}_{h-L:h-1}, \overline{o}_{h-L+1:h-1}, \overline{o}_{N,h}) &= \mathbb{P}_h(\overline{s}_h = \cdot \,|\, \overline{s}_{h-L} \sim \text{Unif}(\mathcal{S}), \overline{a}_{h-L:h-1}, \overline{o}_{h-L+1:h-1}, \overline{o}_{N,h}).
\end{aligned}$$

Then, we have the following lemma.

Lemma .22. There is a constant $C \geq 1$ such that the following holds. Given any LTC problem $\mathcal{L}$ satisfying Assumption III.1, and let $\mathcal{D}'_{\mathcal{L}}$ be the Dec-POMDP after reformulation, strict expansion and refinement. Let $\epsilon \geq 0$, fix a strategy $\overline{g}_{1:\overline{H}}$ and indices $1 \leq h - L < h - 1 \leq \overline{H}$. If $L \geq C\gamma^{-4}\log(\frac{S}{\epsilon})$, then the following set of inequalities hold

$$\mathbb{E}_{\overline{a}_{1:h-1},\overline{o}_{1:h}\sim\overline{g}_{1:\overline{H}}}||\overline{\boldsymbol{b}}_h(\overline{a}_{1:h-1},\overline{o}_{1:h}) - \overline{\boldsymbol{b}}'_h(\overline{a}_{h-L:h-1},\overline{o}_{h-L+1:h})||_1 \leq \epsilon \tag{.16}$$

$$\mathbb{E}_{\overline{a}_{1:h-1},\overline{o}_{1:h}\sim\overline{g}_{1:\overline{H}}}||\overline{\boldsymbol{b}}_h(\overline{a}_{1:h-1},\overline{o}_{1:h-1}) - \overline{\boldsymbol{b}}'_h(\overline{a}_{h-L:h-1},\overline{o}_{h-L+1:h-1})||_1 \leq \epsilon \tag{.17}$$

$$\mathbb{E}_{\overline{a}_{1:h-1},\overline{o}_{1:h}\sim\overline{g}_{1:\overline{H}}}||\overline{\boldsymbol{b}}_h(\overline{a}_{1:h-1},\overline{o}_{1:h-1},\overline{o}_{N,h}) - \overline{\boldsymbol{b}}'_h(\overline{a}_{h-L:h-1},\overline{o}_{h-L+1:h-1},\overline{o}_{N,h})||_1 \leq \epsilon. \tag{.18}$$

Proof. Given any LTC problem $\mathcal{L}$, we can construct a Dec-POMDP $\check{\mathcal{D}}$ that the transition and observation functions of $\check{\mathcal{D}}$ are the same as $\mathcal{L}$. And the information of $\check{\mathcal{D}}$ is fully sharing, which means it shares all the $o_{1:h-1}, a_{1:h}$ as common information at timestep h. Since $\mathcal{D}'_{\mathcal{L}}$ is reformulated from $\mathcal{L}$, we have

$$\overline{\boldsymbol{b}}_h(\overline{a}_{1:h-1},\overline{o}_{1:h}) = \boldsymbol{b}_{\lfloor\frac{h+1}{2}\rfloor}(a_{1:\lfloor\frac{h-1}{2}\rfloor}, o_{1:\lfloor\frac{h+1}{2}\rfloor}) = \check{\boldsymbol{b}}_{\lfloor\frac{h+1}{2}\rfloor}(\check{a}_{1:\lfloor\frac{h-1}{2}\rfloor}, \check{o}_{1:\lfloor\frac{h+1}{2}\rfloor})$$
$$\overline{\boldsymbol{b}}_h(\overline{a}_{1:h-1},\overline{o}_{1:h-1}) = \boldsymbol{b}_{\lfloor\frac{h+1}{2}\rfloor}(a_{1:\lfloor\frac{h-1}{2}\rfloor}, o_{1:\lfloor\frac{h}{2}\rfloor}) = \check{\boldsymbol{b}}_{\lfloor\frac{h+1}{2}\rfloor}(\check{a}_{1:\lfloor\frac{h-1}{2}\rfloor}, \check{o}_{1:\lfloor\frac{h}{2}\rfloor}).$$

And for the approximate belief state, we have

$$\overline{b}'_{h+1}(\overline{a}_{h-L:h},\overline{o}_{h-L+1:h}) = \boldsymbol{b}'_{\lfloor\frac{h+2}{2}\rfloor}(a_{\lfloor\frac{h-L}{2}\rfloor:\lfloor\frac{h}{2}\rfloor}, o_{\lfloor\frac{h-L+2}{2}\rfloor:\lfloor\frac{h+1}{2}\rfloor}) = \check{\boldsymbol{b}}'_{\lfloor\frac{h+2}{2}\rfloor}(\check{a}_{\lfloor\frac{h-L}{2}\rfloor:\lfloor\frac{h}{2}\rfloor}, \check{o}_{\lfloor\frac{h-L+2}{2}\rfloor:\lfloor\frac{h+1}{2}\rfloor})$$
$$\overline{b}'_h(\overline{a}_{h-L:h-1},\overline{o}_{h-L+1:h}) = \boldsymbol{b}'_{\lfloor\frac{h+1}{2}\rfloor}(a_{\lfloor\frac{h-L}{2}\rfloor:\lfloor\frac{h-1}{2}\rfloor}, o_{\lfloor\frac{h-L+2}{2}\rfloor:\lfloor\frac{h+1}{2}\rfloor}) = \check{\boldsymbol{b}}'_{\lfloor\frac{h+1}{2}\rfloor}(\check{a}_{\lfloor\frac{h-L}{2}\rfloor:\lfloor\frac{h-1}{2}\rfloor}, \check{o}_{\lfloor\frac{h-L+2}{2}\rfloor:\lfloor\frac{h}{2}\rfloor}).$$

Also, since for any $t \in [H]$, $\overline{a}_{2t-1}$ are communication actions, $\overline{o}_{2t} = \emptyset$ is null, and $\overline{s}_{2t-1} = \overline{s}_{2t}$ always holds. Then we can write Eq. (.16) and Eq. (.17) as

$$\mathbb{E}_{\{\overline{a}_{2t}\}_{t=1}^{\lfloor\frac{h-1}{2}\rfloor},\{\overline{o}_{2t-1}\}_{t=1}^{\lfloor\frac{h+1}{2}\rfloor}\sim\overline{g}_{1:\overline{H}}}||\overline{\boldsymbol{b}}_h(\overline{a}_{1:h-1},\overline{o}_{1:h}) - \overline{\boldsymbol{b}}'_h(\overline{a}_{h-L:h-1},\overline{o}_{h-L+1:h})||_1 \leq \epsilon \tag{.19}$$

$$\mathbb{E}_{\{\overline{a}_{2t}\}_{t=1}^{\lfloor\frac{h-1}{2}\rfloor},\{\overline{o}_{2t-1}\}_{t=1}^{\lfloor\frac{h+1}{2}\rfloor}\sim\overline{g}_{1:\overline{H}}}||\overline{\boldsymbol{b}}_h(\overline{a}_{1:h-1},\overline{o}_{1:h-1}) - \overline{\boldsymbol{b}}'_h(\overline{a}_{h-L:h-1},\overline{o}_{h-L+1:h-1})||_1 \leq \epsilon. \tag{.20}$$

Since $\check{\mathcal{D}}$ has a fully-sharing IS, for any $i \in [n], h \in [\overline{H}]$ and information $\overline{\tau}_{i,h}, \overline{\tau}_{i,2h}$, we have $\sigma(\overline{\tau}_{i,h}) \subseteq \sigma(\check{\tau}_{i,\lfloor\frac{h+1}{2}\rfloor})$. Therefore, given any strategy $\overline{g}_{1:\overline{H}}$, we can construct a strategy $\check{g}_{1:H}$ such that, for any $\overline{a}_{1:h-1}, \overline{o}_{1:h}$

$$\mathbb{P}(\{\overline{a}_{2t}\}_{t=1}^{\lfloor\frac{h-1}{2}\rfloor}, \{\overline{o}_{2t-1}\}_{t=1}^{\lfloor\frac{h+1}{2}\rfloor} \,|\, \overline{g}_{1:\overline{H}}) = \mathbb{P}(\check{a}_{1:\lfloor\frac{h-1}{2}\rfloor}, \check{o}_{1:\lfloor\frac{h+1}{2}\rfloor} \,|\, \check{g}_{1:H}).$$

Since $\check{\mathcal{D}}$ satisfies Assumption III.1, we can apply the Theorem 10 in [14] with $\check{g}_{1:H}$ to get the result that there is a constant $C_0 \geq 1$ such that if $L' \geq C_0\gamma^{-4}\log(\frac{S}{\epsilon})$, the following holds

$$\mathbb{E}_{\check{a}_{1:\lfloor\frac{h-1}{2}\rfloor},\check{o}_{1:\lfloor\frac{h+1}{2}\rfloor}\sim\check{g}_{1:H}}||\check{\boldsymbol{b}}_{\lfloor\frac{h+1}{2}\rfloor}(\check{a}_{1:\lfloor\frac{h-1}{2}\rfloor}, \check{o}_{1:\lfloor\frac{h+1}{2}\rfloor}) - \check{\boldsymbol{b}}'_{\lfloor\frac{h+1}{2}\rfloor}(\check{a}_{\lfloor\frac{h}{2}\rfloor-L':\lfloor\frac{h-1}{2}\rfloor}, \check{o}_{\lfloor\frac{h+1}{2}\rfloor-L'+1:\lfloor\frac{h+1}{2}\rfloor})||_1 \leq \epsilon \tag{.21}$$

$$\mathbb{E}_{\check{a}_{1:\lfloor\frac{h-1}{2}\rfloor},\check{o}_{1:\lfloor\frac{h+1}{2}\rfloor}\sim\check{g}_{1:H}}||\check{\boldsymbol{b}}_{\lfloor\frac{h+1}{2}\rfloor}(\check{a}_{1:\lfloor\frac{h-1}{2}\rfloor}, \check{o}_{1:\lfloor\frac{h}{2}\rfloor}) - \check{\boldsymbol{b}}'_{\lfloor\frac{h+1}{2}\rfloor}(\check{a}_{\lfloor\frac{h}{2}\rfloor-L':\lfloor\frac{h-1}{2}\rfloor}, \check{o}_{\lfloor\frac{h+1}{2}\rfloor-L'+1:\lfloor\frac{h}{2}\rfloor})||_1 \leq \epsilon. \tag{.22}$$

We choose $C = 3C_0, L = 2L' + 1$. If $L \geq C\gamma^{-4}\log(\frac{S}{\epsilon})$, there must have $L' \geq C_0\gamma^{-4}\log(\frac{S}{\epsilon})$. Therefore, we directly get Eq. (.19) and Eq. (.20).

For Eq. (.18), we cannot directly apply Theorem 10 in [14], but we can slightly change Eq. (E.11) of Theorem 10 in [14] as

$$\mathbb{E}^{\mathcal{D}'_{\mathcal{L}}}_{a_{1:h-1},o_{1:h}\sim g_{1:\overline{H}}}||\overline{\boldsymbol{b}}_h(a_{1:h-1},o_{1:h-1},o_{N,h}) - \overline{\boldsymbol{b}}'_h(a_{h-L:h-1},o_{h-L+1:h-1},o_{N,h})||_1 \leq \epsilon. \tag{.23}$$

It still holds if the posterior update $F^q(P : o_{1,h})$ is changed to $F^q(P : o_{N,h})$, when applying Lemma 9 in the proof of Theorem 10 of [14]. Therefore, we can use the same arguments to prove Eq. (.18) from Eq. (.23) as above, and this completes the proof. □

Then we can compress the common information using a finite-memory truncation. Here, we discuss case-by-case how to compress it for the 8 examples of QC LTC given in §A. Note that after reformulation, strict expansion, and refinement, **Examples 5** and **6** will be the same as **Example 1**, and **Examples 7** and **8** will be the same as **Example 2**. Therefore, we can categorize the examples in §A into 4 types.

Type 1: Baseline sharing of $\mathcal{L}$ is one of **Examples 1, 5, 6** in §A. Then, common information should be that for any $t \in [H], \overline{c}_{2t-1} = \{\overline{o}_{1:2t-2}, \overline{a}_{1:2t-2}\}, \overline{c}_{2t} = \{\overline{o}_{1:2t-2}, \overline{a}_{1:2t-1}, \overline{o}_{N,2t-1}\}, N \subseteq [n]$, where N is the set of agents choose to share their observations through additional sharing, and N can be inferred from $\overline{c}_{2t}$. Then we have that $\mathbb{P}^{\mathcal{D}'_{\mathcal{L}}}_{2t-1}(\overline{s}_{2t-1}, \overline{p}_{2t-1} \,|\, \overline{c}_{2t-1}) = \overline{\boldsymbol{b}}_{2t-1}(\overline{a}_{1:2t-2}, \overline{o}_{1:2t-2})(\overline{s}_{2t-1})\overline{\mathbb{O}}_{2t-1}(\overline{o}_{2t-1} \,|\, \overline{s}_{2t-1})$. Fix compress length $L > 0$, we define the approximate common information as $\widehat{c}_{2t-1} = \{\overline{a}_{2t-1-L:2t-2}, \overline{o}_{2t-L:2t-2}\}$, and the common information conditioned belief as $\mathbb{P}^{\mathcal{M},c}_{2t-1}(\overline{s}_{2t-1}, \overline{p}_{2t-1} \,|\, \widehat{c}_{2t-1}) = \overline{\boldsymbol{b}}_{2t-1}(\overline{a}_{2t-1-L:2t-2}, \overline{o}_{2t-L:2t-2})(\overline{s}_{2t-1})\overline{\mathbb{O}}_{2t-1}(\overline{o}_{2t-1} \,|\, \overline{s}_{2t-1})$. Also, we have $\mathbb{P}^{\mathcal{D}'_{\mathcal{L}}}_{2t}(\overline{s}_{2t}, \overline{p}_{2t} \,|\, \overline{c}_{2t}) =$

$\overline{\boldsymbol{b}}_{2t-1}(\overline{a}_{1:2t-1}, \overline{o}_{1:2t-2}, \overline{o}_{N,2t-1})(\overline{s}_{2t-1})\mathbb{P}_{2t-1}(\overline{o}_{-N,2t-1} \,|\, \overline{s}_{2t-1}, \overline{o}_{N,2t-1})$. Fix compress length $L > 0$, we define the approximate common information a $\widehat{c}_{2t} = \{\overline{a}_{2t-1-L:2t-2}, \overline{o}_{2t-L:2t-2}, \overline{o}_{N,2t-1}\}$, and the common information conditioned belief as $\mathbb{P}^{\mathcal{M},c}_{2t}(\overline{s}_{2t}, \overline{p}_{2t} \,|\, \widehat{c}_{2t}) = \overline{\boldsymbol{b}}'_{2t-1}(\overline{a}_{2t-1-L:2t-2}, \overline{o}_{2t-L:2t-2}, \overline{o}_{N,2t-1})(\overline{s}_{2t-1})\mathbb{P}_{2t-1}(\overline{o}_{-N,2t-1} \,|\, \overline{s}_{2t-1}, \overline{o}_{N,2t-1})$, where $\mathbb{P}_{2t-1}(\overline{o}_{-N,2t-1} \,|\, \overline{s}_{2t-1}, \overline{o}_{N,2t-1}) = \frac{\overline{\mathbb{O}}_{2t-1}(\overline{o}_{N,2t-1}, \overline{o}_{-N,2t-1} \,|\, \overline{s}_{2t-1})}{\sum_{\overline{o}'_{-N,2t-1}} \overline{\mathbb{O}}_{2t-1}(\overline{o}_{N,2t-1}, \overline{o}'_{-N,2t-1} \,|\, \overline{s}_{2t-1})}$. Now, we need to verify that Definition .13 is satisfied.

- The $\{\widehat{c}_h\}_{h\in[\overline{H}]}$ satisfied Eq. (.7) since for any $h \in [H]$, $\widehat{c}_{h+1} \subseteq \widehat{c}_h \cup \overline{z}_h$.
- Note that for any $\overline{c}_{2t-1}$ and the corresponding $\widehat{c}_{2t-1}$ constructed above:

$$
\begin{aligned}
&||\mathbb{P}^{\mathcal{D}'_{\mathcal{L}}}_{2t-1}(\cdot, \cdot \,|\, \overline{c}_h) - \mathbb{P}^{\mathcal{M},c}_{2t-1}(\cdot, \cdot \,|\, \widehat{c}_h)||_1 \\
&= \sum_{\overline{s}_{2t-1}, \overline{o}_{2t-1}} |\overline{\boldsymbol{b}}_{2t-1}(\overline{a}_{1:2t-2}, \overline{o}_{1:2t-2})(\overline{s}_{2t-1})\overline{\mathbb{O}}_{2t-1}(\overline{o}_{2t-1} \,|\, \overline{s}_{2t-1}) \\
&\qquad - \overline{\boldsymbol{b}}'_{2t-1}(\overline{a}_{2t-1-L:2t-2}, \overline{o}_{2t-L:2t-1})(\overline{s}_{2t-1})\overline{\mathbb{O}}_{2t-1}(\overline{o}_{2t-1} \,|\, \overline{s}_{2t-1})| \\
&= ||\overline{\boldsymbol{b}}_{2t-1}(\overline{a}_{1:2t-2}, \overline{o}_{1:2t-2}) - \overline{\boldsymbol{b}}'_{2t-1}(\overline{a}_{2t-1-L:2t-2}, \overline{o}_{2t-L:2t-1})||_1.
\end{aligned}
$$

  For any $\overline{c}_{2t}$ and the corresponding $\widehat{c}_{2t}$ constructed above:

$$
\begin{aligned}
&||\mathbb{P}^{\mathcal{D}'_{\mathcal{L}}}_{2t}(\cdot, \cdot \,|\, \overline{c}_h) - \mathbb{P}^{\mathcal{M},c}_{2t}(\cdot, \cdot \,|\, \widehat{c}_h)|| \\
&= \sum_{\overline{s}_{2t-1}, \overline{o}_{-N,2t-1}} |\overline{\boldsymbol{b}}_{2t-1}(\overline{a}_{1:2t-1}, \overline{o}_{1:2t-2}, \overline{o}_{N,2t-1})(\overline{s}_{2t-1})\mathbb{P}_{2t-1}(\overline{o}_{-N,2t-1} \,|\, \overline{s}_{2t-1}, \overline{o}_{N,2t-1}) \\
&\qquad - \overline{\boldsymbol{b}}'_{2t-1}(\overline{a}_{2t-1-L:2t-2}, \overline{o}_{2t-L:2t-2}, \overline{o}_{N,2t-1})(\overline{s}_{2t-1})\mathbb{P}_{2t-1}(\overline{o}_{-N,2t-1} \,|\, \overline{s}_{2t-1}, \overline{o}_{N,2t-1})| \\
&= ||\overline{\boldsymbol{b}}_{2t-1}(\overline{a}_{1:2t-1}, \overline{o}_{1:2t-2}, \overline{o}_{N,2t-1}) - \overline{\boldsymbol{b}}'_{2t-1}(\overline{a}_{2t-1-L:2t-2}, \overline{o}_{2t-L:2t-2}, \overline{o}_{N,2t-1})||_1.
\end{aligned}
$$

  If we choose $L \geq C\gamma^{-4}\log(\frac{S}{\epsilon})$, then we have that for any $h \in [\overline{H}]$

$$
\mathbb{E}_{\overline{a}_{1:h-1}, o_{1:h} \sim \overline{g}_{1:\overline{H}}} ||\mathbb{P}^{\mathcal{D}'_{\mathcal{L}}}_{h}(\cdot, \cdot \,|\, \overline{c}_h) - \mathbb{P}^{\mathcal{M},c}_{h}(\cdot, \cdot \,|\, \widehat{c}_h)||_1 \leq \epsilon.
$$

Therefore, such a model is an ϵ-expected-approximate common information model.

Type 2: Baseline sharing of $\mathcal{L}$ is **Example 3** in §A. Then, common information common information should be that for any $t \in [H], \overline{c}_{2t-1} = \{\overline{o}_{1:2t-2}, \overline{a}_{1:2t-2}, \overline{o}_{1:2t-1}\}, \overline{c}_{2t} = \{\overline{o}_{1:2t-2}, \overline{a}_{1:2t-1}, \overline{o}_{N,2t-1}\}, N \subseteq [n], 1 \in N$. Here N is the same as defined in case 1, but it must satisfy that $1 \in N$. Then we similarly as case 1, we construct $\widehat{c}_{2t-1} = \{\overline{o}_{2t-L:2t-2}, \overline{a}_{2t-L-1:2t-2}, \overline{o}_{1:2t-1}\}, \widehat{c}_{2t} = \{\overline{a}_{2t-1-L:2t-2}, \overline{o}_{2t-L:2t-2}, \overline{o}_{N,2t-1}\}$, and approximate common information conditioned belief as $\mathbb{P}^{\mathcal{M},c}_{2t-1}(\overline{s}_{2t-1}, \overline{p}_{2t-1} \,|\, \widehat{c}_{2t-1}) = \overline{\boldsymbol{b}}_{2t-1}(\overline{a}_{2t-1-L:2t-2}, \overline{o}_{2t-L:2t-2}, \overline{o}_{1,2t-1})(\overline{s}_{2t-1})\mathbb{P}_{2t-1}(\overline{o}_{-1,2t-1} \,|\, \overline{s}_{2t-1}, \overline{o}_{1,2t-1}), \mathbb{P}^{\mathcal{M},c}_{2t}(\overline{s}_{2t}, \overline{p}_{2t} \,|\, \widehat{c}_{2t}) = \overline{\boldsymbol{b}}'_{2t-1}(\overline{a}_{2t-1-L:2t-2}, \overline{o}_{2t-L:2t-2}, \overline{o}_{N,2t-1})(\overline{s}_{2t-1})\mathbb{P}_{2t-1}(\overline{o}_{-N,2t-1} \,|\, \overline{s}_{2t-1}, \overline{o}_{N,2t-1})$. Now, we need to verify Definition .13 is satisfied.

- The $\{\widehat{c}_h\}_{h\in[\overline{H}]}$ satisfies Eq. (.7) since for any $h \in [H]$, $\widehat{c}_{h+1} \subseteq \widehat{c}_h \cup \overline{z}_h$.
- Note that for any $\overline{c}_{2t-1}$ and the corresponding $\widehat{c}_{2t-1}$ constructed above:

$$
\begin{aligned}
&||\mathbb{P}^{\mathcal{D}'_{\mathcal{L}}}_{2t-1}(\cdot, \cdot \,|\, \overline{c}_h) - \mathbb{P}^{\mathcal{M},c}_{2t-1}(\cdot, \cdot \,|\, \widehat{c}_h)||_1 \\
&= \sum_{\overline{s}_{2t-1}, \overline{o}_{-1,2t-1}} |\overline{\boldsymbol{b}}_{2t-1}(\overline{a}_{1:2t-1}, \overline{o}_{1:2t-2}, \overline{o}_{1,2t-1})(\overline{s}_{2t-1})\mathbb{P}_{2t-1}(\overline{o}_{-1,2t-1} \,|\, \overline{s}_{2t-1}, \overline{o}_{1,2t-1}) \\
&\qquad - \overline{\boldsymbol{b}}'_{2t-1}(\overline{a}_{2t-1-L:2t-2}, \overline{o}_{2t-L:2t-2}, \overline{o}_{1,2t-1})(\overline{s}_{2t-1})\mathbb{P}_{2t-1}(\overline{o}_{-1,2t-1} \,|\, \overline{s}_{2t-1}, \overline{o}_{1,2t-1})| \\
&= ||\overline{\boldsymbol{b}}_{2t-1}(\overline{a}_{1:2t-1}, \overline{o}_{1:2t-2}, \overline{o}_{1,2t-1}) - \overline{\boldsymbol{b}}'_{2t-1}(\overline{a}_{2t-1-L:2t-2}, \overline{o}_{2t-L:2t-2}, \overline{o}_{1,2t-1})||_1.
\end{aligned}
$$

  For any $\overline{c}_{2t}$ and the corresponding $\widehat{c}_{2t}$ constructed above:

$$
\begin{aligned}
&||\mathbb{P}^{\mathcal{D}'_{\mathcal{L}}}_{2t}(\cdot, \cdot \,|\, \overline{c}_h) - \mathbb{P}^{\mathcal{M},c}_{2t}(\cdot, \cdot \,|\, \widehat{c}_h)||_1 \\
&= \sum_{\overline{s}_{2t-1}, \overline{o}_{-N,2t-1}} |\overline{\boldsymbol{b}}_{2t-1}(\overline{a}_{1:2t-1}, \overline{o}_{1:2t-2}, \overline{o}_{N,2t-1})(\overline{s}_{2t-1})\mathbb{P}_{2t-1}(\overline{o}_{-N,2t-1} \,|\, \overline{s}_{2t-1}, \overline{o}_{N,2t-1}) \\
&\qquad - \overline{\boldsymbol{b}}'_{2t-1}(\overline{a}_{2t-1-L:2t-2}, \overline{o}_{2t-L:2t-2}, \overline{o}_{N,2t-1})(\overline{s}_{2t-1})\mathbb{P}_{2t-1}(\overline{o}_{-N,2t-1} \,|\, \overline{s}_{2t-1}, \overline{o}_{N,2t-1})| \\
&= ||\overline{\boldsymbol{b}}_{2t-1}(\overline{a}_{1:2t-1}, \overline{o}_{1:2t-2}, \overline{o}_{N,2t-1}) - \overline{\boldsymbol{b}}'_{2t-1}(\overline{a}_{2t-1-L:2t-2}, \overline{o}_{2t-L:2t-2}, \overline{o}_{N,2t-1})||_1.
\end{aligned}
$$

If we choose $L \geq C\gamma^{-4}\log(\frac{S}{\epsilon})$, then from Lemma .22 we have, for any $h \in [\overline{H}]$

$$\mathbb{E}_{\overline{a}_{1:h-1}, o_{1:h} \sim \overline{g}_{1:\overline{H}}} ||\mathbb{P}_h^{\mathcal{D}'_{\mathcal{L}}}(\cdot, \cdot \,|\, \overline{c}_h) - \mathbb{P}_h^{\mathcal{M}, c}(\cdot, \cdot \,|\, \widehat{c}_h)||_1 \leq \epsilon.$$

Therefore, such a model is an ϵ-expected-approximate common information model.

Type 3: Baseline sharing of $\mathcal{L}$ is one of **Examples 2, 7, 8** in §A. Then the common information should be that, for any $h \in [\overline{H}], \overline{c}_h = \{\overline{o}_{1:h-2d}, \overline{a}_{1,1:h-1}, \{\overline{a}_{-1,2t-1}\}_{t=\lfloor \frac{h-2d+1}{2} \rfloor}^{\lfloor \frac{h}{2} \rfloor}, \overline{o}_{1,h-2d+1:h}, \overline{o}_M\}$, where $M \subset \{(i,t) \,|\, 1 < i \leq n, h-2d+1 \leq t \leq h\}$ and $\overline{o}_M = \{o_{i,t} \,|\, (i,t) \in M\}$, and corresponding $\overline{p}_h = \{\overline{o}_{i,t} \,|\, 1 < i \leq n, h-2d < t \leq h, (i,t) \notin M\}$. Actually, $\overline{o}_M$ are the observations shared by the additional sharing in $\mathcal{L}$. Denote $f_{\tau,h-2d} = \{\overline{a}_{1:h-2d-1}, \overline{o}_{h-2d}, \{\overline{a}_{-1,2t-1}\}_{t=\lfloor \frac{h-2d+1}{2} \rfloor}^{\lfloor \frac{h}{2} \rfloor}\}, f_a = \{\overline{a}_{1,h-2d:h-1}\}, f_o = \{\overline{o}_{1,h-2d+1:h}, \overline{o}_M\}$. We can compute the common-information-based belief as

$$\begin{aligned}
\mathbb{P}_h^{\mathcal{D}'_{\mathcal{L}}}(\overline{s}_h, \overline{p}_h \,|\, \overline{c}_h) &= \sum_{\overline{s}_{h-2d}} \mathbb{P}_h^{\mathcal{D}'_{\mathcal{L}}}(\overline{s}_h, \overline{p}_h \,|\, \overline{s}_{h-2d}, f_a, f_o) \mathbb{P}_h^{\mathcal{D}'_{\mathcal{L}}}(\overline{s}_{h-2d} \,|\, f_{\tau,h-2d}, f_a, f_o) \\
&= \sum_{\overline{s}_{h-2d}} \mathbb{P}_h^{\mathcal{D}'_{\mathcal{L}}}(\overline{s}_h, \overline{p}_h \,|\, \overline{s}_{h-2d}, f_a, f_o) \frac{\mathbb{P}_h^{\mathcal{D}'_{\mathcal{L}}}(\overline{s}_{h-2d}, f_a, f_o \,|\, f_{\tau,h-2d})}{\sum_{\overline{s}'_{h-2d}} \mathbb{P}_h^{\mathcal{D}'_{\mathcal{L}}}(\overline{s}'_{h-2d}, f_a, f_o \,|\, f_{\tau,h-2d})}.
\end{aligned}$$

Denote the probability $P_h(f_o \,|\, \overline{s}_{h-2d}, f_a) := \Pi_{t=1}^{2d} \mathbb{P}_h^{\mathcal{D}'_{\mathcal{L}}}(\overline{o}_{1,h-2d+t}, \overline{o}_{M_{h-2d+t}} \,|\, \overline{s}_{h-2d}, \overline{a}_{1,h-2d:h-2d+t})$, where $M_{h-2d+t} = \{(i, h-2d+t) \,|\, (i, h-2d+t) \in M\}$ denotes the set of observations at timestep $h-2d+t$ and shared through additional sharing. With such notation, we have

$$\begin{aligned}
\mathbb{P}_h^{\mathcal{D}'_{\mathcal{L}}}(\overline{s}_{h-2d} \,|\, f_{\tau,h-2d}, f_a, f_o) =& \frac{\overline{\boldsymbol{b}}_{h-2d}(\overline{a}_{1:h-2d-1}, \overline{o}_{1:h-2d})(\overline{s}_{h-2d}) P_h(f_o \,|\, \overline{s}_{h-2d}, f_a)}{\sum_{\overline{s}'_{h-2d}} \overline{\boldsymbol{b}}_{h-2d}(\overline{a}_{1:h-2d-1}, \overline{o}_{1:h-2d})(\overline{s}'_{h-2d}) P_h(f_o \,|\, \overline{s}'_{h-2d}, f_a)} \\
=& F^{P_h(\cdot \,|\, \cdot, f_a)}(\overline{\boldsymbol{b}}_{h-2d}(\overline{a}_{1:h-2d-1}, \overline{o}_{1:h-2d}); f_o)(\overline{s}_{h-2d}),
\end{aligned}$$

where $F^{P_h(\cdot \,|\, \cdot, f_a)}(\cdot; f_o) : \Delta(\mathcal{S}) \to \Delta(\mathcal{S})$ is the posterior belief update function. The formal definition is shown in Lemma 9 in [14].

Then, we define the approximate common information as $\widehat{c}_h := \{\overline{o}_{1,h-2d-L+1:h}, \overline{a}_{1,h-2d-L:h-1}, \overline{o}_M\}$ and corresponding approximate common information conditioned belief as

$$\mathbb{P}_h^{\mathcal{M}, c}(\overline{s}_h, \overline{p}_h \,|\, \widehat{c}_h) = \sum_{\overline{s}_{h-2d}} \mathbb{P}_h^{\mathcal{D}'_{\mathcal{L}}}(\overline{s}_h, \overline{p}_h \,|\, \overline{s}_{h-2d}, f_a, f_o) F^{P_h(\cdot \,|\, \cdot, f_a)}(\overline{\boldsymbol{b}}'_{h-2d}(\overline{a}_{h-2d-L:h-2d-1}, \overline{o}_{h-2d-L+1:h-2d}); f_o)(\overline{s}_{h-2d}).$$

Now we verify that Definition .13 is satisfied.

- Obviously, the $\{\widehat{c}_h\}_{h \in [\overline{H}}$ satisfies Eq. (.7).
- For any $\overline{c}_h$ and the corresponding $\widehat{c}_h$ constructed above:

$$\begin{aligned}
&||\mathbb{P}_h^{\mathcal{D}'_{\mathcal{L}}}(\cdot, \cdot \,|\, \overline{c}_h) - \mathbb{P}_h^{\mathcal{M}, c}(\cdot, \cdot \,|\, \widehat{c}_h)||_1 \\
&\quad \leq ||F^{P(\cdot \,|\, \cdot, f_a)}(\overline{\boldsymbol{b}}_{h-2d}(\overline{a}_{1:h-2d-1}, \overline{o}_{1:h-2d}); f_o) - F^{P(\cdot \,|\, \cdot, f_a)}(\overline{\boldsymbol{b}}'_{h-2d}(\overline{a}_{h-2d-L:h-2d-1}, \overline{o}_{h-2d-L+1:h-2d}); f_o)||_1.
\end{aligned}$$

  If we choose $L \geq C\gamma^{-4}\log(\frac{S}{\epsilon})$, then for any strategy $\overline{g}_{1:\overline{H}}$, by taking expectations over $f_{\tau,h-2d}, f_a, f_o$, from Lemma .22 and Lemma 9 in [14], we have, for any $h \in [\overline{H}]$

$$\mathbb{E}_{\overline{a}_{1:h-1}, o_{1:h} \sim \overline{g}_{1:\overline{H}}} ||\mathbb{P}_h^{\mathcal{D}'_{\mathcal{L}}}(\cdot, \cdot \,|\, \overline{c}_h) - \mathbb{P}_h^{\mathcal{M}, c}(\cdot, \cdot \,|\, \widehat{c}_h)||_1 \leq \epsilon.$$

Therefore, such a model is an ϵ-expected-approximate common information model.

Type 4: Baseline sharing of $\mathcal{L}$ is **Example 4** in §A. Then, for any $h \in [H]$, the common information should be $\widehat{c}_h = \{\overline{o}_{1:h-2d}, \{\overline{a}_{2t-1}\}_{t=1}^{\lfloor \frac{h}{2} \rfloor}, \overline{o}_M\}$, where $M = \{(i,t) \,|\, i \in [n], h-2d+1 \leq t \leq h\}$. Then, still we denote $f_{\tau,h-2d} = \{\overline{o}_{1:h-2d}, \{\overline{a}_{2t-1}\}_{t=1}^{\lfloor \frac{h}{2} \rfloor}\}, f_o = \{\overline{o}_M\}$. We can compute the common-information-based belief as

$$\begin{aligned}
\mathbb{P}_h^{\mathcal{D}'_{\mathcal{L}}}(\overline{s}_h, \overline{p}_h \,|\, \overline{c}_h) &= \sum_{\overline{s}_{h-2d}} \mathbb{P}_h^{\mathcal{D}'_{\mathcal{L}}}(\overline{s}_h, \overline{p}_h \,|\, \overline{s}_{h-2d}, f_o) \mathbb{P}_h^{\mathcal{D}'_{\mathcal{L}}}(\overline{s}_{h-2d} \,|\, f_{\tau,h-2d}, f_o) \\
&= \sum_{\overline{s}_{h-2d}} \mathbb{P}_h^{\mathcal{D}'_{\mathcal{L}}}(\overline{s}_h, \overline{p}_h \,|\, \overline{s}_{h-2d}, f_o) \frac{\mathbb{P}_h^{\mathcal{D}'_{\mathcal{L}}}(\overline{s}_{h-2d}, f_o \,|\, f_{\tau,h-2d})}{\sum_{\overline{s}'_{h-2d}} \mathbb{P}_h^{\mathcal{D}'_{\mathcal{L}}}(\overline{s}'_{h-2d}, f_o \,|\, f_{\tau,h-2d})}.
\end{aligned}$$

Denote the probability $P_h(f_o \mid \overline{s}_{h-2d}) := \Pi_{t=1}^{2d} \mathbb{P}_h^{\mathcal{D}'_{\mathcal{L}}}(\overline{o}_{1,h-2d+t}, \overline{o}_{M_{h-2d+t}} \mid \overline{s}_{h-2d})$, where $M_{h-2d+t} = \{(i, h-2d+t) \mid (i, h-2d+t) \in M\}$ denotes the set of observations at timestep $h-2d+t$ and shared through additional sharing. Since the actions do not influence underlying states, here we use the belief notation $\overline{\boldsymbol{b}}_k(\overline{o}_{1:k}), \overline{\boldsymbol{b}}_k(\overline{o}_{k-L:k})$, $\forall k \in [\overline{H}], L < k$. With such notation, we have

$$\mathbb{P}_h^{\mathcal{D}'_{\mathcal{L}}}(\overline{s}_{h-2d} \mid f_{\tau,h-2d}, f_o) = \frac{\overline{\boldsymbol{b}}_{h-2d}(\overline{o}_{1:h-2d})(\overline{s}_{h-2d}) P_h(f_o \mid \overline{s}_{h-2d})}{\sum_{\overline{s}'_{h-2d}} \overline{\boldsymbol{b}}_{h-2d}(\overline{o}_{1:h-2d})(\overline{s}'_{h-2d}) P_h(f_o \mid \overline{s}'_{h-2d})} = F^{P_h(\cdot \mid \cdot)}(\overline{\boldsymbol{b}}_{h-2d}(\overline{o}_{1:h-2d}); f_o)(\overline{s}_{h-2d}),$$

where $F^{P_h(\cdot \mid \cdot)}(\cdot; f_o) : \Delta(\mathcal{S}) \to \Delta(\mathcal{S})$ is the posterior belief update function, the same as mentioned in **Type 3**.
Then, we define the approximate common information as $\widehat{c}_h := \{\overline{o}_{h-2d-L+1:h}, \overline{o}_M\}$ and corresponding approximate common information conditioned belief as

$$\mathbb{P}_h^{\mathcal{M},c}(\overline{s}_h, \overline{p}_h \mid \widehat{c}_h) = \sum_{\overline{s}_{h-2d}} \mathbb{P}_h^{\mathcal{D}'_{\mathcal{L}}}(\overline{s}_h, \overline{p}_h \mid \overline{s}_{h-2d}, f_o) F^{P_h(\cdot \mid \cdot)}(\overline{\boldsymbol{b}}'_{h-2d}(\overline{o}_{h-2d-L+1:h-2d}); f_o)(\overline{s}_{h-2d}).$$

Now we verify that Definition .13 is satisfied.

- Obviously, the $\{\widehat{c}_h\}_{h \in [\overline{H}]}$ satisfies Eq.(.7).
- For any $\overline{c}_h$ and corresponding $\widehat{c}_h$ constructed above:

$$\begin{aligned} &||\mathbb{P}_h^{\mathcal{D}'_{\mathcal{L}}}(\cdot, \cdot \mid \overline{c}_h) - \mathbb{P}_h^{\mathcal{M},c}(\cdot, \cdot \mid \widehat{c}_h)||_1 \\ &\quad \le ||F^{P(\cdot \mid \cdot)}(\overline{\boldsymbol{b}}_{h-2d}(\overline{o}_{1:h-2d}); f_o) - F^{P(\cdot \mid \cdot)}(\overline{\boldsymbol{b}}'_{h-2d}(\overline{a}_{h-2d-L:h-2d-1}, \overline{o}_{h-2d-L+1:h-2d}); f_o)||_1. \end{aligned}$$

  If we choose $L \ge C\gamma^{-4} \log(\frac{S}{\epsilon})$, then for any strategy $\overline{g}_{1:\overline{H}}$, by taking expectations over $f_{\tau,h-2d}, f_o$, from Lemma .22 and Lemma 9 in [14], we have, for any $h \in [\overline{H}]$

$$\mathbb{E}_{\overline{a}_{1:h-1}, o_{1:h} \sim \overline{g}_{1:\overline{H}}} ||\mathbb{P}_h^{\mathcal{D}'_{\mathcal{L}}}(\cdot, \cdot \mid \overline{c}_h) - \mathbb{P}_h^{\mathcal{M},c}(\cdot, \cdot \mid \widehat{c}_h)||_1 \le \epsilon.$$

Therefore, such a model is an ϵ-expected-approximate common information model.

Combining **Parts I, II, III**, we complete the proof. □

Remark .23. Let $\mathcal{L}$ be an LTC problem satisfying Assumptions III.1, III.4, III.5, and III.7, and $\mathcal{D}'_{\mathcal{L}}$ be the Dec-POMDP after reformulation, strict expansion and refinement. Then, if $\mathcal{L}$ has any one of baseline sharing protocols as in §A, and $\mathcal{L}$ satisfies the conditions as follows, then $\mathcal{D}'_{\mathcal{L}}$ satisfies Assumption IV.7.

- If $\mathcal{L}$ has baseline sharing protocol as one of **Examples 1, 5, 6** in §A, $\mathcal{L}$ needs to satisfy the **Part (1)** of **Factorized structure** in §G.
- If $\mathcal{L}$ has baseline sharing protocol as one of **Examples 2, 3, 4, 7, 8** in §A, it does not need additional condition.
- For the examples with a delay d of sharing, i.e., **Examples 2, 4, 7, 8**, our main result in Theorem .18 on the quasi-polynomial time-complexity also applies to the case when $d = \text{poly}\log H$ (beyond being a *constant* as assumed throughout), since the total complexity will contain error bound of order $\mathcal{O}((|\mathcal{O}_h||\mathcal{A}_h|)^d)$. A similar generalization also applies to the main result for learning in LTCs (see Theorem IV.9). Such a generalization also resembles that on the quasi-polynomial complexities for the examples in Theorems 7 and 9 in [14].

Actually, such a condition is also considered in [14]. For $\mathcal{L}$ with baseline sharing protocols as one of the examples in §A and satisfying the conditions as above, we can construct the expected common information model $\mathcal{M}$ of $\mathcal{D}'_{\mathcal{L}}$ as mentioned in the proof of Theorem .18. If the baseline sharing protocol of $\mathcal{L}$ is one of **Examples 1, 5, 6**, then $\mathcal{D}'_{\mathcal{L}}$ and $\mathcal{M}$ satisfy **Factorized structures** condition in §G; If the baseline sharing protocol of $\mathcal{L}$ is one of **Examples 2, 7, 8**, then $\mathcal{D}'_{\mathcal{L}}$ and $\mathcal{M}$ satisfy **Turn-based structures** condition in §G; If the baseline sharing protocol of $\mathcal{L}$ is one of **Examples 3, 4**, then $\mathcal{D}'_{\mathcal{L}}$ and $\mathcal{M}$ satisfy **Nested private information** condition in §G. From Lemma .27, we can conclude that Assumption IV.7 holds.

8) Main Results for Learning in QC LTC:

Proof of Theorem IV.9.
Firstly, given any LTC problem $\mathcal{L}$, we can construct $\mathcal{D}'_{\mathcal{L}}$ by doing reformulation, strict expansion and refinement from $\mathcal{L}$. According to Theorem IV.5, we know that $\mathcal{D}'_{\mathcal{L}}$ has SI-CIBs w.r.t. strategy spaces $\overline{\Gamma}_{1:\overline{H}}$.
Secondly, given any LTC problem $\mathcal{L}$, we can apply Algorithm 2 to solve such problem. From the proof of .18, we know that Algorithm 6 can output the team optimal strategy of $\widehat{\mathcal{M}}(\overline{g}^{1:\overline{H},j})$ for each $j \in [K]$. Then, from Theorem 4 in [14], it can guarantee that $\overline{g}^*_{1:\overline{H}}$ is an ϵ-team optimum of $\mathcal{D}'_{\mathcal{L}}$ with probability at least $1-\delta_1$, where $\epsilon = \overline{H}\epsilon_r(\widetilde{\mathcal{M}}(\overline{g}^{1:\overline{H}})) + \overline{H}^2\epsilon_z(\widetilde{\mathcal{M}}(\overline{g}^{1:\overline{H}})) + (\overline{H}^2 + \overline{H})\epsilon_{apx}(\overline{g}^{1:\overline{H}}, \widehat{L}, \zeta_1, \zeta_2, \theta_1, \theta_2, \phi)$. Then, from the proof of Theorem .18, we have that given $\overline{g}^*_{1:\overline{H}}$ as an ϵ-team optimal strategy of $\mathcal{D}'_{\mathcal{L}}$, we can get $(g^{m,*}_{1:H}, g^{a,*}_{1:H})$ as an ϵ-team optimal strategy of $\mathcal{L}$.
Lastly, for any $\mathcal{L}$ has one of the baseline sharing protocols as in §A and has any one of the structures in §G, then we can

construct $\widehat{M} = \widehat{M}(\overline{g}^{1:\overline{H},j})$ for any $j \in [K]$ by calling Algorithm 5 of [30], and enforces the model to satisfy the structures in §G. Then, such $\widehat{M}$ satisfies Assumption IV.7 according to Lemma .27. Therefore, we can learn a $\overline{g}^*_{1:\overline{H}}$ as an ϵ-team optimum of $\mathcal{D}'_{\mathcal{L}}$ with probability at least $1-\delta_1$, and then get $(g^{m,*}_{1:H}, g^{a,*}_{1:H})$ as an ϵ-team optimal strategy of $\mathcal{L}$. Thus, we complete the proof. □

D. Deferred Details of §V

In the following, we will use $\bar{}$ to denote the elements and random variables in the Dec-POMDP $\mathcal{D}$. We first introduce the notion of *perfect recall* [22]:

Definition .24 (Perfect recall)**.** We say that agent i has perfect recall if $\forall h = 2, \cdots, \overline{H}$, it holds that $\overline{\tau}_{i,h-1} \cup \{\overline{a}_{i,h-1}\} \subseteq \overline{\tau}_{i,h}$. If for any $i \in [n]$, agent i has perfect recall, we call that the Dec-POMDP has a perfect recall property.

1) Proof of Theorem V.1:

Proof. sQC $\Rightarrow$ SI-CIB:
Let $\mathcal{D}$ be the Dec-POMDP with an sQC information structure, and $\mathcal{D}$ satisfy Assumptions II.2, III.5, and III.7. To prove that $\mathcal{D}$ has SI-CIBs, it is sufficient to prove that for any $h = 2, \cdots, \overline{H}$, fix any $h_1 \in [h-1], i_1 \in [n]$, and for any $\overline{g}_{1:h-1} \in \overline{\mathcal{G}}_{1:h-1}, \overline{g}'_{i_1,h_1} \in \overline{\mathcal{G}}_{i_1,h_1}$, let $\overline{g}'_{h_1} := (\overline{g}_{1,h_1}, \cdots, \overline{g}'_{i_1,h_1}, \cdots, \overline{g}_{n,h_1})$ and $\overline{g}'_{1:h-1} := (\overline{g}_1, \cdots, \overline{g}'_{h_1}, \cdots, \overline{g}_{h-1})$, the following holds

$$\mathbb{P}(\overline{s}_h, \overline{p}_h \,|\, \overline{c}_h, \overline{g}_{1:h-1}) = \mathbb{P}(\overline{s}_h, \overline{p}_h \,|\, \overline{c}_h, \overline{g}'_{1:h-1}). \tag{.24}$$

We prove this case-by-case as follows:

1) If there exists some $i_3 \neq i_1$ such that $\sigma(\overline{\tau}_{i_1,h_1}) \cup \sigma(\overline{a}_{i_1,h_1}) \subseteq \sigma(\overline{\tau}_{i_3,h})$, then from Assumption II.2, we know that $\sigma(\overline{\tau}_{i_1,h_1}) \cup \sigma(\overline{a}_{i_1,h_1}) \subseteq \sigma(\overline{c}_h)$. Therefore, there exist deterministic functions β_1, β_2 such that $\overline{\tau}_{i_1,h_1} = \beta_1(\overline{c}_h), \overline{a}_{i_1,h_1} = \beta_2(\overline{c}_h)$, and further it holds that

$$\begin{aligned}\mathbb{P}(\overline{s}_h, \overline{p}_h \,|\, \overline{c}_h, \overline{g}_{1:h-1}) &= \mathbb{P}(\overline{s}_h, \overline{p}_h \,|\, \beta_1(\overline{c}_h), \beta_2(\overline{c}_h), \overline{c}_h, \overline{g}_{1:h-1}) \\ &= \mathbb{P}(\overline{s}_h, \overline{p}_h \,|\, \overline{\tau}_{i_1,h_1}, \overline{a}_{i_1,h_1}, \overline{c}_h, \overline{g}_{1:h-1}) = \mathbb{P}(\overline{s}_h, \overline{p}_h \,|\, \overline{\tau}_{i_1,h_1}, \overline{a}_{i_1,h_1}, \overline{c}_h, \overline{g}'_{1:h-1}).\end{aligned}$$

   The last equality is due to the fact that the input and output of $\overline{g}_{i_1,h_1}$ are $\overline{\tau}_{i_1,h_1}$ and $\overline{a}_{i_1,h_1}$, respectively.
2) If there does not exist any $i_2 \neq i_1$ such that $\sigma(\overline{\tau}_{i_1,h_1}) \cup \sigma(\overline{a}_{i_1,h_1}) \subseteq \sigma(\overline{\tau}_{i_2,h})$, i.e., for all $i_2 \neq i_1$, either $\sigma(\overline{\tau}_{i_1,h_1}) \nsubseteq \sigma(\overline{\tau}_{i_2,h})$ or $\sigma(\overline{a}_{i_1,h_1}) \nsubseteq \sigma(\overline{\tau}_{i_2,h})$, then agent (i_1,h_1) does not influence agent (i_2,h) for any $i_2 \neq i_1$, since $\mathcal{D}$ is sQC. Firstly, we claim that agent (i_1,h_1) does not influence $\overline{s}_{h_1+1}$: since if it influences, from Assumption III.7, there exists some $i_3 \neq i_1$ such that agent (i_1,h_1) influences $\overline{o}_{i_3,h_1+1}$; however, from Assumption II.1 (e), we know $\overline{o}_{i_3,h_1+1} \in \overline{\tau}_{i_3,h_1+1} \subseteq \overline{\tau}_{i_3,h}$; therefore, agent (i_1,h_1) influences agent (i_3,h), contradicting the argument above that the former does not influence (i_2,h) for any $i_2 \neq i_1$. Hence, we further have that agent (i_1,h_1) does not influence $\overline{s}_{h_2}$ for any $h_2 > h_1$. Therefore, by Assumption III.5, for any $h_2 > h_1$, $\overline{a}_{i_1,h_1} \notin \overline{\tau}_{h_2}$.
   Secondly, we claim that agent (i_1,h_1) does not influence $\overline{\tau}_{i_4,h_2}$, for any $i_4 \in [n]$ and $h_2 > h_1$. This is because of the fact that agent (i_1,h_1) does not influence $\overline{s}_{h_1+1}$ and thus not $\overline{o}_{i_4,h_1+1}$ for any $i_4 \in [n]$, together with the fact proved above that $\overline{a}_{i_1,h_1} \notin \overline{\tau}_{h_1+1}$, implies that agent (i_1,h_1) does not influence any element in $\overline{\tau}_{i_4,h_1+1}$ for any $i_4 \in [n]$, either directly or indirectly. Since $\overline{\tau}_{i_4,h_1+1}$ is the input of agent i_4's strategy at timestep h_1+1 to decide $\overline{a}_{i_4,h_1+1}$, agent (i_1,h_1) thus does not influence $\overline{a}_{i_4,h_1+1}$ for any $i_4 \in [n]$, either, which, together with the fact that it does not influence $\overline{s}_{h_1+2}$ and thus not $\overline{o}_{i_4,h_1+2}$ for any $i_4 \in [n]$, further implies that it does not influence any element in $\overline{\tau}_{i_4,h_1+2}$ for any $i_4 \in [n]$. By recursion, agent (i_1,h_1) does not influence $\overline{\tau}_{i_4,h_2}$ for any $i_4 \in [n]$ and $h_2 > h_1$.
   Therefore, agent (i_1,h_1) does not influence $\overline{c}_h = \cap_{i_4=1}^n \overline{\tau}_{i_4,h}$ nor $\overline{p}_h = \overline{\tau}_h \backslash \overline{c}_h$, which thus leads to

$$\mathbb{P}(\overline{s}_h, \overline{p}_h \,|\, \overline{c}_h, \overline{g}_{1:h-1}) = \mathbb{P}(\overline{s}_h, \overline{p}_h \,|\, \overline{c}_h, \overline{g}'_{1:h-1}).$$

SI-CIB $\Rightarrow$ sQC:
Since $\mathcal{D}$ has perfect recall and has SI-CIBs, i.e., $\forall i \in [n], h \in [\overline{H}], \forall \overline{g}_{1:h-1}, \overline{g}'_{1:h-1} \in \overline{\mathcal{G}}_{1:h-1}, \overline{c}_h \in \overline{\mathcal{C}}_h, \overline{s}_h \in \overline{\mathcal{S}}, \overline{p}_h \in \overline{\mathcal{P}}_h$, the following holds

$$\mathbb{P}(\overline{s}_h, \overline{p}_h \,|\, \overline{c}_h, \overline{g}_{1:h-1}) = \mathbb{P}(\overline{s}_h, \overline{p}_h \,|\, \overline{c}_h, \overline{g}'_{1:h-1}).$$

Our goal is to prove that $\mathcal{D}$ is sQC (up to null sets). In particular, we meant to prove that if agent (i_1,h_1) influences agent (i_2,h_2) in the intrinsic model of the Dec-POMDP (see §F), then under any strategy $\overline{g}_{1:\overline{H}} \in \overline{\mathcal{G}}_{1:\overline{H}}$, $\sigma(\overline{\tau}_{i_1,h_1}) \subseteq \sigma(\overline{\tau}_{i_2,h_2})$ except the null sets generated by $\overline{g}_{1:\overline{H}}$.

We prove this by contradiction. If this is not true, then there exists some strategy $\overline{g}_{1:\overline{H}}$ and $i_1, i_2 \in [n], h_1, h_2 \in [\overline{H}]$, such that agent (i_1,h_1) influences agent (i_2,h_2), but either $\sigma(\overline{\tau}_{i_1,h_1}) \nsubseteq \sigma(\overline{\tau}_{i_2,h_2})$ or $\sigma(\overline{a}_{i_1,h_1}) \nsubseteq \sigma(\overline{\tau}_{i_2,h_2})$ (up to the null sets generated by $\overline{g}_{1:\overline{H}}$). First, we can assume $i_2 \neq i_1$, since otherwise it always holds that $\overline{\tau}_{i_1,h_1} \subseteq \overline{\tau}_{i_1,h_2}$ and $\overline{a}_{i_1,h_1} \in \overline{\tau}_{i_1,h_2}$, due to the assumption that the agents in $\mathcal{D}$ have perfect recall.

Then, we discuss the following different cases. Note that in the following discussion, when it comes to σ-algebra inclusion, we meant it up to the null sets generated by $\overline{g}_{1:\overline{H}}$.

1) If $\sigma(\overline{a}_{i_1,h_1}) \nsubseteq \sigma(\overline{\tau}_{i_2,h_2})$, then it implies that $\sigma(\overline{a}_{i_1,h_1}) \nsubseteq \sigma(\overline{c}_{h_2})$ because $\overline{c}_{h_2} \subseteq \overline{\tau}_{i_2,h_2}$. This also implies that $\overline{a}_{i_1,h_1} \notin \overline{c}_{h_2}$, and thus $\overline{a}_{i_1,h_1} \in \overline{p}_{i_1,h_2}$ due to perfect recall. Note that there must exist some realizations $\overline{c}_{h_2} \in \overline{\mathcal{C}}_{h_2}, \overline{p}_{h_2} \in \overline{\mathcal{P}}_{h_2}, \overline{s}_{h_2} \in \overline{\mathcal{S}}$ such that $\overline{c}_{h_2}$ has non-zero probability under $\overline{g}_{1:h_2-1}$, and $\mathbb{P}(\overline{s}_{h_2}, \overline{p}_{h_2} \,|\, \overline{c}_{h_2}, \overline{g}_{1:h_2-1}) \neq 0$. Meanwhile, there must exist another different action realization $\overline{a}'_{i_1,h_1}$ such that

$$\mathbb{P}(\overline{s}_{h_2}, \overline{p}_{h_2} \backslash \{\overline{a}_{i_1,h_1}\} \cup \{\overline{a}'_{i_1,h_1}\} \,|\, \overline{c}_{h_2}, \overline{g}_{1:h_2-1}) \neq 0, \tag{.25}$$

since otherwise it holds that $\sigma(\overline{a}_{i_1,h_1}) \subseteq \sigma(\overline{c}_{h_2})$. Actually, this means that there are some non-zero probability trajectories containing $\overline{a}_{i_1,h_1}$ and $\overline{c}_{h_2}$, and some non-zero probability trajectories containing $\overline{a}'_{i_1,h_1}$ and $\overline{c}_{h_2}$. Then, we define another strategy $\overline{g}'_{i_1,h_1}$ as:

$$\forall \overline{\tau}_{i_1,h_1} \in \overline{\mathcal{T}}_{i_1,h_1}, \quad \overline{g}'_{i_1,h_1}(\overline{\tau}_{i_1,h_1}) = \overline{a}'_{i_1,h_1}, \tag{.26}$$

and we let $\overline{g}'_{h_1} := (\overline{g}_{1,h_1}, \cdots, \overline{g}'_{i_1,h_1}, \cdots, \overline{g}_{n,h_1})$ and $\overline{g}'_{1:h_2-1} := (\overline{g}_1, \cdots, \overline{g}'_{h_1}, \cdots, \overline{g}_{h_2-1})$.
Now we claim that $\overline{c}_{h_2}$ has non-zero probability under $\overline{g}'_{1:h_2-1}$. From that $\overline{c}_{h_2}$ has non-zero probability under $\overline{g}_{1:h_2-1}$, and $\mathbb{P}(\overline{s}_{h_2}, \overline{p}_{h_2} \backslash \{\overline{a}_{i_1,h_1}\} \cup \{\overline{a}'_{i_1,h_1}\} \,|\, \overline{c}_{h_2}, \overline{g}_{1:h_2-1}) \neq 0$, we can get $\mathbb{P}(\overline{a}'_{i_1,h_1}, \overline{c}_{h_2} \,|\, \overline{g}_{1:h_2-1}) > 0$. Since $\overline{g}'_{1:h_2-1}$ only differs from $\overline{g}_{1:h_2-1}$ in the strategy of agent (i_1, h_1), and $\overline{g}'_{i_1,h_1}$ always chooses $\overline{a}'_{i_1,h_1}$, then we get $\mathbb{P}(\overline{a}'_{i_1,h_1}, \overline{c}_{h_2} \,|\, \overline{g}'_{1:h_2-1}) \geq \mathbb{P}(\overline{a}'_{i_1,h_1}, \overline{c}_{h_2} \,|\, \overline{g}_{1:h_2-1}) > 0$ because $\overline{g}_{1:h_2-1}$ and $\overline{g}'_{1:h_2-1}$ are the same in those trajectories containing $\overline{a}'_{i_1,h_1}$ and $\overline{c}_{h_2}$, and thus $\mathbb{P}(\overline{c}_{h_2} \,|\, \overline{g}'_{1:h_2-1}) > 0$. Hence, we prove our claim.
Meanwhile, due to (.26), notice that

$$\mathbb{P}(\overline{s}_{h_2}, \overline{p}_{h_2} \,|\, \overline{c}_{h_2}, \overline{g}'_{1:h_2-1}) = 0 \neq \mathbb{P}(\overline{s}_{h_2}, \overline{p}_{h_2} \,|\, \overline{c}_{h_2}, \overline{g}_{1:h_2-1}), \tag{.27}$$

which leads to a contradiction to the fact that $\mathcal{D}$ has SI-CIBs.

2) If $\sigma(\overline{a}_{i_1,h_1}) \subseteq \sigma(\overline{\tau}_{i_2,h_2})$, then it implies that $\sigma(\overline{\tau}_{i_1,h_1}) \nsubseteq \sigma(\overline{\tau}_{i_2,h_2})$, and further implies that $\sigma(\overline{\tau}_{i_1,h_1}) \nsubseteq \sigma(\overline{c}_{h_2})$ since $\overline{c}_{h_2} \subseteq \overline{\tau}_{i_2,h_2}$. Note that there must exist some realizations $\overline{c}_{h_2} \in \overline{\mathcal{C}}_{h_2}, \overline{\tau}_{i_2,h_2} \in \overline{\mathcal{T}}_{i_2,h_2}$ such that $\overline{\tau}_{i_2,h_2}$ has non-zero probability under $\overline{g}_{1:h_2-1}$ and $\overline{c}_{h_2} \subseteq \overline{\tau}_{i_2,h_2}$, and there exist two realizations $\overline{\tau}_{i_1,h_1}, \overline{\tau}'_{i_1,h_1} \in \overline{\mathcal{T}}_{i_1,h_1}$ such that $\mathbb{P}(\overline{\tau}_{i_1,h_1} \,|\, \overline{\tau}_{i_2,h_2}) > 0, \mathbb{P}(\overline{\tau}'_{i_1,h_1} \,|\, \overline{\tau}_{i_2,h_2}) > 0$, since otherwise, it holds that $\sigma(\overline{\tau}_{i_1,h_1}) \subseteq \sigma(\overline{c}_{h_2})$. Furthermore, we know that there exist $\overline{s}_{h_2} \in \overline{\mathcal{S}}, \overline{p}_{h_2} \in \overline{\mathcal{P}}_{h_2}$ such that $\mathbb{P}(\overline{s}_{h_2}, \overline{p}_{h_2} \,|\, \overline{c}_{h_2}, \overline{g}_{1:h_2-1}) > 0$ and $\overline{\tau}'_{i_2,h_2} \subseteq \overline{c}_{h_2} \cup \overline{p}_{h_2}$. Since $\sigma(\overline{a}_{i_1,h_1}) \subseteq \sigma(\overline{\tau}_{i_2,h_2})$, we know that there exists $\overline{a}_{i_1,h_1}$ that $\mathbb{P}(\overline{a}_{i_1,h_1} \,|\, \overline{\tau}_{i_2,h_2}) = 1$. Let $\tau := \overline{\tau}_{i_1,h_1} \backslash \overline{c}_{h_2}$ and $\tau' := \overline{\tau}'_{i_1,h_1} \backslash \overline{c}_{h_2}$. and consider another action $\overline{a}'_{i_1,h_1} \neq \overline{a}_{i_1,h_1}$ and strategy $\overline{g}'_{i_1,h_1}$ defined such that

$$\overline{g}'_{i_1,h_1}(\overline{\tau}_{i_1,h_1}) = \overline{a}'_{i_1,h_1}, \quad \overline{g}'_{i_1,h_1}(\overline{\tau}'_{i_1,h_1}) = \overline{a}_{i_1,h_1}, \tag{.28}$$

and keeps $g'_{i_1,h_1}(\overline{\tau}''_{i_1,h_1})$ the same as $g_{i_1,h_1}(\overline{\tau}''_{i_1,h_1})$ for any other $\overline{\tau}''_{i_1,h_1}$. We denote $\overline{g}'_{h_1} := (\overline{g}_{1,h_1}, \cdots, \overline{g}'_{i_1,h_1}, \cdots, \overline{g}_{n,h_1})$ and $\overline{g}'_{1:h_2-1} := (\overline{g}_1, \cdots, \overline{g}'_{h_1}, \cdots, \overline{g}_{h_2-1})$. Since $(\overline{\tau}'_{i_1,h_1}, \overline{\tau}_{i_2,h_2})$ has non-zero probability under $\overline{g}_{1:h_2-1}$ and $\mathbb{P}(\overline{a}_{i_1,h_1} \,|\, \overline{\tau}_{i_2,h_2})$, then we know $(\overline{\tau}'_{i_1,h_1}, \overline{\tau}_{i_2,h_2})$ has non-zero probability under $\overline{g}_{;1:h_2-1}$. Hence, we know that $\overline{c}_{h_2}$ has non-zero probability under $\overline{g}_{;1:h_2-1}$. Meanwhile, it holds that

$$\mathbb{P}(\overline{s}_{h_2}, \overline{p}_{h_2} \,|\, \overline{c}_{h_2}, \overline{g}'_{1:h_2-1}) = \frac{\mathbb{P}(\overline{s}_{h_2}, \overline{p}_{h_2}, \overline{c}_{h_2} \,|\, \overline{g}'_{1:h_2-1})}{\mathbb{P}(\overline{c}_{h_2} \,|\, \overline{g}'_{1:h_2-1})} = \frac{\mathbb{P}(\overline{s}_{h_2}, \overline{\tau}_{h_2} \,|\, \overline{g}'_{1:h_2-1})}{\mathbb{P}(\overline{c}_{h_2} \,|\, \overline{g}'_{1:h_2-1})} = 0 \neq \mathbb{P}(\overline{s}_{h_2}, \overline{p}_{h_2} \,|\, \overline{c}_{h_2}, \overline{g}_{1:h_2-1}), \tag{.29}$$

where the third equal sign is because $\overline{a}_{i_1,h_1} \in \overline{\tau}_{h_2}, \overline{\tau}_{i_1,h_1} \subseteq \overline{\tau}_{h_2}$ from perfect recall, and $\overline{a}_{i_1,h_1}, \overline{\tau}_{i_1,h_1}$ can never happen simultaneously under $\overline{g}'_{1:h_2-1}$ due to (.28). Therefore, (.29) leads to a contradiction to the fact that $\mathcal{D}$ has SI-CIBs.

This completes the proof. □

E. Collection of Algorithm Pseudocodes

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

F. Decentralized POMDPs (with Information Sharing)

A Dec-POMDP with n agents and potential information sharing can be characterized by a tuple

$$\mathcal{D} = \langle H, \mathcal{S}, \{\mathcal{A}_{i,h}\}_{i\in[n],h\in[H]}, \{\mathcal{O}_{i,h}\}_{i\in[n],h\in[H]}, \{\mathbb{T}_h\}_{h\in[H]}, \{\mathbb{O}_h\}_{h\in[H]}, \mu_1, \{\mathcal{R}_h\}_{h\in[H]} \rangle,$$

where H denotes the length of each episode, $\mathcal{S}$ denotes state space, and $\mathcal{A}_{i,h}$ denotes the *control action* spaces of agent i at timestep h. We denote by $s_h \in \mathcal{S}$ the state and by $a_{i,h}$ the control action of agent i at timestep h. We use $a_h := (a_{1,h}, \cdots, a_{n,h}) \in \mathcal{A}_h := \mathcal{A}_{1,h} \times \mathcal{A}_{2,h} \times \cdots \mathcal{A}_{n,h}$ to denote the joint control action for all the n agents at timestep h, with $\mathcal{A}_h$ denoting the joint control action space at timestep h. We denote $\mathbb{T} = \{\mathbb{T}_h\}_{h\in[H]}$ the collection of transition functions,

[2]We here slightly abuse the notation by using $\widetilde{\tau}_{i_1,h_1}, \widetilde{c}_{h_2}$ to denote both the random variables and their realizations. Note that the inclusion of σ-algebras here does not rely on the realized values of $\widetilde{\tau}_{i_1,h_1}, \widetilde{c}_{h_2}$, but relies on the information structure of $\mathcal{D}_{\mathcal{L}}$.

Algorithm 1 Planning in QC LTC Problems

Require: LTC $\mathcal{L}$, accuracy levels $\epsilon_r, \epsilon_z > 0$
Reformulate $\mathcal{L}$ to $\mathcal{D}_{\mathcal{L}}$ based on Eq. (IV.1)
Expand $\mathcal{D}_{\mathcal{L}}$ to $\mathcal{D}^{\dagger}_{\mathcal{L}}$ based on Eq. (IV.2)
Refine $\mathcal{D}^{\dagger}_{\mathcal{L}}$ to $\mathcal{D}'_{\mathcal{L}}$ based on $\mathcal{L}$ and Eq. (IV.4)
Construct expected Approximate Common-information Model $\mathcal{M}$ from $\mathcal{D}'_{\mathcal{L}}$ with error ϵ_r, ϵ_z
$\overline{g}^*_{1:\widetilde{H}} \leftarrow$ Algorithm 6($\mathcal{M}$)
$\widetilde{g}^*_{1:\widetilde{H}} \leftarrow \varphi(\overline{g}^*_{1:\widetilde{H}}, \mathcal{D}_{\mathcal{L}})$
$g^{m,*}_{1:H} \leftarrow \{\widetilde{g}^*_1, \widetilde{g}^*_3, \cdots, \widetilde{g}^*_{2H-1}\}$
$g^{a,*}_{1:H} \leftarrow \{\widetilde{g}_2, \widetilde{g}_4, \cdots, \widetilde{g}_{2H}\}$
return $(g^{m,*}_{1:H}, g^{a,*}_{1:H})$

Algorithm 2 Learning in QC LTC Problems

Require: Underlying environment LTC $\mathcal{L}$, iteration number K
Reformulate $\mathcal{L}$ to $\mathcal{D}_{\mathcal{L}}$ based on Eq. (IV.1)
Refine $\mathcal{D}_{\mathcal{L}}$ to $\mathcal{D}'_{\mathcal{L}}$ based on Eq. (IV.2)
Obtain $\{\overline{g}^{1:\overline{H},j}\}_{j=1}^{K}$ by calling Algorithm 3 of [31]
for $j = 1$ to K **do**
  Construct $\widehat{\mathcal{M}}(\overline{g}^{1:\overline{H},j})$ by calling Algorithm 5 of [14] with the underlying environment $\mathcal{D}'_{\mathcal{L}}$ and $\overline{g}^{1:\overline{H},j}$
  $\overline{g}^{j,*}_{1:\overline{H}} \leftarrow$ Algorithm 6($\widehat{\mathcal{M}}(\overline{g}^{1:\overline{H},j})$)
end for
$\overline{g}^*_{1:\overline{H}} \leftarrow$ Algorithm 8($\{\overline{g}^{j,*}_{1:\overline{H}}\}_{j=1}^{K}$) of [14]
$\widetilde{g}^*_{1:\widetilde{H}} \leftarrow \varphi(\overline{g}^*_{1:\overline{H}}, \mathcal{D}_{\mathcal{L}})$
$g^{m,*}_{1:H} \leftarrow \{\widetilde{g}^*_1, \widetilde{g}^*_3, \cdots, \widetilde{g}^*_{2H-1}\}$
$g^{a,*}_{1:H} \leftarrow \{\widetilde{g}_2, \widetilde{g}_4, \cdots, \widetilde{g}_{2H}\}$
return $(g^{m,*}_{1:H}, g^{a,*}_{1:H})$

Algorithm 3 Vanilla Realization of $\varphi(\breve{g}_{1:\breve{H}}, \mathcal{D}_{\mathcal{L}})$

Require: Strategy $\breve{g}_{1:\breve{H}}$, QC Dec-POMDP $\mathcal{D}_{\mathcal{L}}$
$\widetilde{g}_{1:\widetilde{H}} \leftarrow \emptyset$
for $h_2 = 1$ to $\breve{H}$, $i_2 = 1$ to n, $\widetilde{\tau}_{i_2,h_2} \in \widetilde{\mathcal{T}}_{i_2,h_2}$ **do**
  $\breve{\tau}_{i_2,h_2} \leftarrow \widetilde{\tau}_{i_2,h_2}$
  for $h_1 = 1$ to $h_2 - 1$, $i_1 = 1$ to n **do**
    if $\sigma(\widetilde{\tau}_{i_1,h_1}) \subseteq \sigma(\widetilde{c}_{h_2})$[2] in $\mathcal{D}_{\mathcal{L}}$ **then**
      Obtain the value of $\widetilde{\tau}_{i_1,h_1}$ from that of $\widetilde{c}_{h_2}$ (based on $\widetilde{\tau}_{i_2,h_2}$)
      $\widetilde{a}_{i_1,h_1} \leftarrow \widetilde{g}_{i_1,h_1}(\widetilde{\tau}_{i_1,h_1})$
      $\breve{\tau}_{i_2,h_2} \leftarrow \breve{\tau}_{i_2,h_2} \cup \{\widetilde{a}_{i_1,h_1}\}$
    end if
  end for
  $\widetilde{g}_{i_2,h_2}(\widetilde{\tau}_{i_2,h_2}) \leftarrow \breve{g}_{i_2,h_2}(\breve{\tau}_{i_2,h_2})$
end for
return $\widetilde{g}_{1:\widetilde{H}}$

Algorithm 4 Efficient Implementation of $\varphi(\breve{g}_{1:\breve{H}}, \mathcal{D}_{\mathcal{L}})$

Require: Strategy $\breve{g}_{1:\breve{H}}$, QC Dec-POMDP $\mathcal{D}_{\mathcal{L}}$
for $h = 1$ to $\breve{H}$ **do**
  for $i = 1$ to n **do**
    Agent i receives $\widetilde{\tau}_{i,h}$
    $\breve{\tau}_{i,h} \leftarrow$ Recover($\widetilde{\tau}_{i,h}, \breve{g}_{1:h-1}, \mathcal{D}_{\mathcal{L}}$) \\ Recursion of Algorithm 5
    Agent i takes action $\widetilde{a}_{i,h} \leftarrow \breve{g}_{i,h}(\breve{\tau}_{i,h})$
  end for
end for

Algorithm 5 Recover($\widetilde{\tau}_{i,h}, \breve{g}_{1:h-1}, \mathcal{D}_{\mathcal{L}}$)

Require: Information $\widetilde{\tau}_{i,h}$, Strategy $\breve{g}_{1:h-1}$, QC Dec-POMDP $\mathcal{D}_{\mathcal{L}}$
  $\breve{\tau}_{i,h} \leftarrow \widetilde{\tau}_{i,h}$
  for $j = 1$ to n, $h' = 1$ to $h-1$ **do**
    if $\sigma(\widetilde{\tau}_{j,h'}) \subseteq \sigma(\widetilde{c}_h)$ in $\mathcal{D}_{\mathcal{L}}$ and $\widetilde{a}_{j,h'} \notin \widetilde{\tau}_{i,h}$ **then**
      Obtain the value of $\widetilde{\tau}_{j,h'}$ from that of $\widetilde{c}_h$ (based on $\widetilde{\tau}_{i,h}$)
      $\breve{\tau}_{j,h'} \leftarrow$ Recover$(\widetilde{\tau}_{j,h'}, \breve{g}_{1:h'-1}, \mathcal{D}_{\mathcal{L}})$
      $\widetilde{a}_{j,h'} \leftarrow \breve{g}_{j,h'}(\breve{\tau}_{j,h'})$
      $\breve{\tau}_{i,h} \leftarrow \breve{\tau}_{i,h} \cup \{\widetilde{a}_{j,h'}\}$
    end if
  end for
  return $\breve{\tau}_{i,h}$

Algorithm 6 Planning in Dec-POMDP with expected Approximate Common-information Model

Require: Expected Approximate Common-information Model $\mathcal{M}$.
  for $i \in [n]$ and $\widehat{c}_{\overline{H}+1} \in \widehat{\mathcal{C}}_{\overline{H}+1}$ **do**
    $V^{*,\mathcal{M}}_{i,\overline{H}+1}(\widehat{c}_{\overline{H}+1}) \leftarrow 0$
  end for
  for $h = \overline{H}$ to 1 **do**
    for $\widehat{c}_h \in \widehat{\mathcal{C}}_h$ **do**
      Define $Q^{*,\mathcal{M}}_h(\widehat{c}_h, \gamma_{1,h}, \cdots, \gamma_{n,h}) := \widehat{\mathcal{R}}^{\mathcal{M}}_h(\widehat{c}_h, \gamma_h) + \mathbb{E}^{\mathcal{M}}\left[V^{*,\mathcal{M}}_{h+1}(\widehat{c}_{h+1}) \,|\, \widehat{c}_h, \gamma_h\right]$

$$\left(\widehat{g}^*_{1,h}(\cdot \,|\, \widehat{c}_h, \cdot), \cdots, \widehat{g}^*_{n,h}(\cdot \,|\, \widehat{c}_h, \cdot)\right) \leftarrow \underset{\gamma_{1:n,h} \in \Gamma_h}{\operatorname{argmax}} Q^{*,\mathcal{M}}_h(\widehat{c}_h, \gamma_{1,h}, \cdots, \gamma_{n,h}) \tag{.30}$$

    end for
    $V^{*,\mathcal{M}}_h(\widehat{c}_h) \leftarrow \max_{\gamma_{1:n,h}} Q^{*,\mathcal{M}}_h(\widehat{c}_h, \gamma_{1,h}, \cdots, \gamma_{n,h})$
  end for
  return $\widehat{g}^*_{1:\overline{H}}$

where $\mathbb{T}_h(\cdot \,|\, s_h, a_h) \in \Delta(\mathcal{S})$ gives the transition probability to the next state s_{h+1} when taking the joint control action a_h at state s_h. We use $\mu_1 \in \Delta(\mathcal{S})$ to denote the distribution of the initial state s_1. We denote by $\mathcal{O}_{i,h}$ the observation space and by $o_{i,h} \in \mathcal{O}_{i,h}$ the observation of agent i at timestep h. We use $o_h := (o_{1,h}, o_{2,h}, \cdots, o_{n,h}) \in \mathcal{O}_h := \mathcal{O}_{1,h} \times \mathcal{O}_{2,h} \times \cdots \mathcal{O}_{n,h}$ to denote the joint observation of all the n agents at timestep h, with $\mathcal{O}_h$ denoting the joint observation space at timestep h. We use $\{\mathbb{O}_h\}_{h\in[H]}$ to denote the collection of emission matrices, where $o_h \sim \mathbb{O}_h(\cdot \,|\, s_h) \in \Delta(\mathcal{O}_h)$ at timestep h under state $s_h \in \mathcal{S}$. For notational convenience, we adopt the matrix convention, where $\mathbb{O}_h$ is a matrix with each row $\mathbb{O}_h(\cdot \,|\, s_h)$ for all $s_h \in \mathcal{S}$. Also, we denote by $\mathbb{O}_{i,h}$ the marginalized emission for agent i at timestep h. Finally, $\{\mathcal{R}_h\}_{h\in[H]}$ is a collection of reward functions among all agents, where $\mathcal{R}_h : \mathcal{S} \times \mathcal{A}_h \to [0,1]$.

At timestep h, each agent i in the Dec-POMDP has access to some information $\tau_{i,h}$, a subset of historical joint observations and actions, namely, $\tau_{i,h} \subseteq \{o_1, a_1, o_2, \cdots, a_{h-1}, o_h\}$, and the collection of all possible such available information is denoted by $\mathcal{T}_{i,h}$. We use τ_h to denote the *joint* available information at timestep h. Meanwhile, agents may *share* part of the history with each other. The *common information* $c_h = \cup_{t=1}^h z_t$ at timestep h is thus a subset of the joint history τ_h, where z_h is the information shared at timestep h. We use $\mathcal{C}_h$ to denote the collection of all possible c_h at timestep h, and use $\mathcal{T}_{i,h}$ to denote the collection of all possible $\tau_{i,h}$ of agent i at timestep h. Besides the common information c_h, each agent also has her *private information* $p_{i,h} = \tau_{i,h} \backslash c_h$, where the collection of $p_{i,h}$ is denoted by $\mathcal{P}_{i,h}$. We also denote by p_h the *joint* private information, and by $\mathcal{P}_h$ the collection of all possible p_h at timestep h. We refer to the above the *state-space model* of the Dec-POMDP (with information sharing).

Each agent i at timestep h chooses the control action $a_{i,h}$ based on some strategy $g_{i,h} : \mathcal{T}_{i,h} \to \mathcal{A}_{i,h}$. We denote by $g_h := (g_{1,h}, g_{2,h}, \cdots, g_{n,h})$ the joint control strategy of all the agents, and by $g_{1:h} := (g_1, g_2, \cdots, g_h), \forall h \in [H]$ the sequence of joint strategies from timestep 1 to h. We use $\mathcal{G}_{i,h}$ to denote the strategy space of $g_{i,h}$, and use $\mathcal{G}_h, \mathcal{G}_{1:h}$ to denote joint strategy spaces, correspondingly.

Next, we introduce some background on the intrinsic model and information structure of Dec-POMDPs.

1) Intrinsic Model: In an intrinsic model [23], we regard the agent i at different timesteps as *different agents*, and each agent only acts *once* throughout. Any Dec-POMDP $\mathcal{D}$ with n agents can be formulated within the intrinsic-model framework, and can be characterized by a tuple $\langle(\Omega, \mathscr{F}), N, \{(\mathbb{U}_l, \mathscr{U}_l)\}_{l=1}^N, \{(\mathbb{I}_l, \mathscr{I}_l)\}_{l=1}^N\rangle$ [11], where $(\Omega, \mathcal{F})$ is a measurable space of

the environment, $N = n \times H$ is the number of agents in the intrinsic model. By a slight abuse of notation, we write $[N] := [n] \times [H]$, and write $l := (i,h) \in [N]$ for notational convenience. This way, any agent $l \in [N]$ corresponds to an agent $i \in [n]$ at timestep $h \in [H]$ in the state-space model. We denote by $\mathbb{U}_l$ the measurable action space of agent l and by $\mathscr{U}_l$ the σ-algebra over $\mathbb{U}_l$. For $A \subseteq [N]$, let $\mathbb{H}_A := \Omega \times \prod_{l \in A} \mathbb{U}_l$ and $\mathbb{H} := \mathbb{H}_{[N]}$. For any σ-algebra $\mathscr{C}$ over $\mathbb{H}_A$, let $\langle \mathscr{C} \rangle$ denote the cylindrical extension of $\mathscr{C}$ on $\mathbb{H}$. Let $\mathscr{H}_A := \langle \mathscr{F} \otimes (\otimes_{l \in A} \mathscr{U}_l) \rangle$ and $\mathscr{H} = \mathscr{H}_{[N]}$. We denote by $\mathbb{I}_l$ the space of *information available* to agent l, and by $\mathscr{I}_l$ the σ-algebra over $\mathbb{H}$. For $l \in [N]$, we denote by I_l the information of agent l, and U_l the action of agent l. The spaces and random variables of agent $l = (i,h)$ in the intrinsic model are related to those in the state-space model as follows: $\forall l = (i,h) \in [N], \mathbb{U}_l = \mathcal{A}_{i,h}, \mathbb{I}_l = \mathcal{T}_{i,h}, U_l = a_{i,h}, I_l = \tau_{i,h}$.

2) Information Structures of Dec-POMDPs: An important class of IS is the *quasi-classical* one, which is defined as follows [23], [11], [12].

Definition .25 (Quasi-classical Dec-POMDPs)**.** We call a Dec-POMDP problem *QC* if each agent in the intrinsic model knows the information available to the agents who influence her, directly or indirectly, i.e. $\forall l_1, l_2 \in [N], l_1 = (i_1, h_1), l_2 = (i_2, h_2), i_1, i_2 \in [n], h_1, h_2 \in [H]$, if agent l_1 influences agent l_2, then $\mathscr{I}_{l_1} \subseteq \mathscr{I}_{l_2}$.

Furthermore, *strictly* quasi-classical IS [23], [24], as a subclass of QC IS, is defined as follows.

Definition .26 (Strictly quasi-classical Dec-POMDPs)**.** We call a Dec-POMDP problem *sQC* if each agent in the intrinsic model knows the information *and* actions available to the agents who influence her, directly or indirectly. That is, $\forall l_1, l_2 \in [N], l_1 = (i_1, h_1), l_2 = (i_2, h_2), i_1, i_2 \in [n], h_1, h_2 \in [H]$, if agent l_1 influences agent l_2, then $\mathscr{I}_{l_1} \cup \langle \mathscr{U}_{l_1} \rangle \subseteq \mathscr{I}_{l_2}$.

3) Intrinsic Model of LTC Problems: Given any LTC $\mathcal{L}$ of the state-space-model form defined in §II-A, we define the intrinsic model of $\mathcal{L}$ as a tuple $\langle (\Omega, \mathscr{F}), N, \{(\mathbb{U}_l, \mathscr{U}_l)\}_{l=1}^N, \{(\mathbb{M}_l, \mathscr{M}_l)\}_{l=1}^N, \{(\mathbb{I}_{l^-}, \mathscr{I}_{l^-})\}_{l=1}^N, \{(\mathbb{I}_{l^+}, \mathscr{I}_{l^+})\}_{l=1}^N \rangle$, where $(\Omega, \mathscr{F})$ is the measure space representing all the uncertainty in the system; $N = n \times H$ is the number of agents in the intrinsic model. By a slight abuse of notation, we write $[N] := [n] \times [H]$, and write $l := (i,h) \in [N]$ for convenience. This way, any agent $l \in [N]$ corresponds to an agent $i \in [n]$ at timestep $h \in [H]$ in the state-space model, and we thus define $l^- := (i, h^-)$ and $l^+ := (i, h^+)$ accordingly. We denote by $\mathbb{U}_l$ and $\mathbb{M}_l$ the measurable control and communication action spaces of agent l, and by $\mathscr{U}_l$ and $\mathscr{M}_l$ the σ-algebra over $\mathbb{U}_l$ and $\mathbb{M}_l$, respectively. For any $A \subseteq [N]$, let $\mathbb{H}_A := \Omega \times \prod_{l \in A} (\mathbb{U}_l \times \mathbb{M}_l)$ and $\mathbb{H} := \mathbb{H}_{[N]}$. For any σ-algebra $\mathscr{C}$ over $\mathbb{H}_A$, let $\langle \mathscr{C} \rangle$ denote the cylindrical extension of $\mathscr{C}$ on $\mathbb{H}$. Let $\mathscr{H}_A := \langle \mathscr{F} \otimes (\otimes_{l \in A} \mathscr{U}_l) \otimes (\otimes_{l \in A} \mathscr{M}_l) \rangle$, $\mathscr{H} = \mathscr{H}_{[N]}$. We denote by $\mathbb{I}_{l^-}$ and $\mathbb{I}_{l^+}$ the spaces of *information available* to agent l *before* and *after* additional sharing, respectively, and by $\mathscr{I}_{l^-} \subseteq \mathscr{H}$ and $\mathscr{I}_{l^+} \subseteq \mathscr{H}$ the associated σ-algebra. The spaces and random variables of agent $l = (i,h)$ in the intrinsic model are related to those in the state-space model as follows: $\forall l = (i,h) \in [N], \mathbb{U}_l = \mathcal{A}_{i,h}, \mathbb{M}_l = \mathcal{M}_{i,h}, \mathbb{I}_{l^-} = \mathcal{T}_{i,h^-}, \mathbb{I}_{l^+} = \mathcal{T}_{i,h^+}, U_l = a_{i,h}, M_l = m_{i,h}, I_{l^-} = \tau_{i,h^-}, I_{l^+} = \tau_{i,h^+}$. For notational convenience, for any random variable B in LTC and the σ-algebra $\mathscr{B}$ generated by B, we overload $\sigma(B)$ to denote the cylindrical extension of $\mathscr{B}$ on $\mathbb{H}$, i.e., $\sigma(B) = \langle \mathscr{B} \rangle$.

G. Conditions Leading to Assumption IV.7

As a minimal requirement for computational tractability (for both Dec-POMDPs and LTCs), Assumption IV.7 is needed for the one-step tractability of the team-decision problem involved in the value iteration in Algorithm 6. We now adapt several such structural conditions from [14] to the LTC setting, which lead to this assumption and have been studied in the literature. Note that since we need to do planning in the approximate model $\mathcal{M}$, which is oftentimes constructed based on the original problem $\mathcal{L}$ and approximate belief $\{\mathbb{P}_h^{\mathcal{M},c}(\overline{s}_h, \overline{p}_h \,|\, \widehat{c}_h)\}_{h \in [\overline{H}]}$, we necessarily need assumptions on these two models $\mathcal{L}$ and $\mathcal{M}$, for which we refer to as the **Part (1)** and **Part (2)** of the conditions below, respectively.

- **Turn-based structures. Part (1):** At each timestep $h \in [H]$, there is only one agent, denoted as $ct(h) \in [n]$, that can affect the state transition. More concretely, the transition dynamics take the forms of $\mathbb{T}_h : \mathcal{S} \times \mathcal{A}_{ct(h)} \to \Delta(\mathcal{S})$. Additionally, we assume the reward function admits an additive structure such that $\mathcal{R}_h(s_h, a_h) = \sum_{i \in [n]} \mathcal{R}_{i,h}(s_h, a_{i,h})$ for some functions $\{\mathcal{R}_{i,h}\}_{i \in [n]}$. Meanwhile, since only agent $ct(h)$ takes the action, we assume the increment of the common information $z_{h+1}^b = \chi_{h+1}(p_{h^+}, a_{ct(h),h}, o_{h+1})$. **Part (2):** No additional requirement. Such a structure has been commonly studied in (fully observable) stochastic games and multi-agent RL [32], [33].
- **Nested private information. Part (1):** No additional requirement. **Part (2):** At each timestep $h = 2t, t \in [H]$, all the agents form a *hierarchy* according to the private information after $a_{i,h}$ they possess, in the sense that $\forall\ i,j \in [n], j < i, \overline{p}_{j,h} = Y_h^{i,j}(\overline{p}_{i,h})$ for some function $Y_h^{i,j}$. More formally, the approximate belief satisfies that $\mathbb{P}_h^{\mathcal{M},c}(\overline{p}_{j,h} = Y_h^{i,j}(\overline{p}_{i,h}) \,|\, \overline{p}_{i,h}, \widehat{c}_h) = 1$. Such a structure has been investigated in [34] with heuristic search, and in [14] with finite-time complexity analysis.
- **Factorized structures. Part (1):** At each timestep $h \in [H]$, the state s_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 $\mathbb{T}_h(s_{h+1} \,|\, s_h, a_h) = \prod_{i=1}^n \mathbb{T}_{i,h}(s_{i,h+1} \,|\, s_{i,h}, a_{i,h})$, the emission also takes the product form of $\mathbb{O}_h(o_h \,|\, s_h) = \prod_{i=1}^n \mathbb{O}_{i,h}(o_{i,h} \,|\, s_{i,h})$, and the communication cost and reward functions can be decoupled into n terms such that $\mathcal{K}_h(p_h, m_h) =$

$\sum_{i=1}^n \mathcal{K}_{i,h}(p_{i,h}, m_{i,h}), \mathcal{R}_h(s_h, a_h) = \sum_{i=1}^n \mathcal{R}_{i,h}(s_{i,h}, a_{i,h})$. **Part (2):** At each timestep $h \in [\overline{H}]$, the approximate common information is also factorized so that $\widehat{c}_h = (\widehat{c}_{1,h}, \widehat{c}_{2,h}, \cdots, \widehat{c}_{n,h})$ and its evolution satisfies that $\widehat{c}_{i,h+1} = \widehat{\phi}_{i,h+1}(\widehat{c}_{i,h}, \overline{z}_{i,h})$ for some function $\widehat{\phi}_{i,h+1}$. Correspondingly, the approximate belief need to satisfy that $\mathbb{P}_h^{\mathcal{M},c}(\overline{s}_h, \overline{p}_h \,|\, \widehat{c}_h) = \Pi_{i=1}^n \mathbb{P}_{i,h}^{\mathcal{M},c}(\overline{s}_{i,h}, \overline{p}_{i,h} \,|\, \widehat{c}_{i,h})$ for some functions $\{\mathbb{P}_{i,h}^{\mathcal{M},c}\}_{i\in[n],h\in[\overline{H}]}$ Such a structure, under general information sharing protocols, can lead to non-classical IS. In this case, it can be viewed an example of non-classical ISs where the agents have no incentive for signaling [12, §3.8.3].

Lemma .27. Given any LTC problem $\mathcal{L}$ and $\mathcal{D}'_{\mathcal{L}}$ is the Dec-POMDP after reformulation, expansion and refinement. For any $\mathcal{M}$ to be the approximate model of $\mathcal{D}'_{\mathcal{L}}$ and $\{\mathbb{P}_h^{\mathcal{M},c}\}_{h\in[\overline{H}]}$ to be the approximate belief, if they satisfy any of the 3 conditions above, then Eq. (.30) in Algorithm 6 can be solved in polynomial time, i.e., Assumption IV.7 holds.

Proof. We prove the result case by case:

- **Nested private information:** For any $h = 2t-1, t \in [H]$, from the Assumption III.4 and construction of $\mathcal{D}'_{\mathcal{L}}$, since we need to find the optimal strategy of $\mathcal{D}'_{\mathcal{L}}$ in the spaces $\overline{g}_h \in \overline{\mathcal{C}}_h \to \overline{\mathcal{A}}_h$, (recall that $\overline{\mathcal{A}}_h = \mathcal{M}_t$ is joint communication action space), we can regard the private information of $\mathcal{D}'_{\mathcal{L}}$ as empty, i.e. $\overline{p}_{i,h} = \emptyset$. Therefore, we know that for any two agents i, j, $\overline{p}_{j,h} = \emptyset = Y_h^{i,j}(\overline{p}_{i,h})$, with $Y_h^{i,j}$ defined as $Y_h^{i,j}(\overline{p}_{i,h}) =$ for any $i, j \in [n], \overline{p}_{i,h} \in \mathcal{P}_{i,h}$. Therefore, for both odd steps $h = 2t-1$ and even steps $h = 2t, t \in [H]$, the private information are nested.
  Then, for any $h \in [\overline{H}]$, we first define the $u_{i,h} \in \mathcal{U}_{i,h} := \{(\times_{j=1}^i \overline{\mathcal{P}}_{j,h}) \times (\times_{j=1}^{i-1} \overline{\mathcal{A}}_{j,h}) \to \overline{\mathcal{A}}_{i,h}\}$ and slightly abuse the notation for $Q_h^{*,\mathcal{M}}$ as follows

$$Q_h^{*,\mathcal{M}}(\widehat{c}_h, u_{1,h}, \cdots, u_{n,h}) := \sum_{\overline{s}_h, \overline{p}_h, \overline{a}_h, \overline{s}_{h+1}, \overline{o}_{h+1}} \mathbb{P}_h^{\mathcal{M},c}(\overline{s}_h, \overline{p}_h \,|\, \widehat{c}_h) \Pi_{i=1}^n \mathbb{1}[\overline{a}_{i,h} = u_{i,h}(\overline{p}_{1:i,h}, \overline{a}_{1:i-1,h})] \overline{\mathbb{T}}_h(\overline{s}_{h+1} \,|\, \overline{s}_h, \overline{a}_h)$$
$$\overline{\mathbb{O}}_{h+1}(\overline{o}_{h+1} \,|\, \overline{s}_{h+1}) [\overline{\mathcal{R}}_h(\overline{s}_h, \overline{a}_h) + V_{h+1}^{*,\mathcal{M}}(\widehat{c}_{h+1})].$$

  Since the space of $\mathcal{U}_{i,h}$ covers the space $\Gamma_{i,h}$, then for the $u_{1:n,h}^*$ be an optimal one that maximize the $Q_h^{*,\mathcal{M}}$, we have

$$Q_h^{*,\mathcal{M}}(\widehat{c}_h, u_{1,h}^*, \cdots, u_{n,h}^*) = \max_{\{u_{i,h} \in \mathcal{U}_{i,h}\}_{i\in[n]}} Q_h^{*,\mathcal{M}}(\widehat{c}_h, u_{1,h}, \cdots, u_{n,h}) \geq \max_{\{\gamma_{i,h} \in \Gamma_{i,h}\}_{i\in[n]}} Q_h^{*,\mathcal{M}}(\widehat{c}_h, \gamma_{1,h}, \cdots, \gamma_{n,h}).$$

  Meanwhile, due to the nested private information condition, for any $\overline{p}_h \in \overline{\mathcal{P}}_h$, there must exists $\gamma'_{1:n,h}$ such that $\gamma'_{1:n,h}$ output the same actions as $u_{1:n,h}^*$ under $\overline{p}_h$. Therefore, we can conclude that

$$\max_{\{u_{i,h} \in \mathcal{U}_{i,h}\}_{i\in[n]}} Q_h^{*,\mathcal{M}}(\widehat{c}_h, u_{1,h}, \cdots, u_{n,h}) = \max_{\{\gamma_{i,h} \in \Gamma_{i,h}\}_{i\in[n]}} Q_h^{*,\mathcal{M}}(\widehat{c}_h, \gamma_{1,h}, \cdots, \gamma_{n,h})$$

  Therefore, we can solve Eq. (.30) and compute $\gamma_{1:n,h}^*$ from computing $u_{1:n,h}^*$, which can be solved with complexity poly$(\overline{\mathcal{P}}_h, \overline{\mathcal{A}}_h, \overline{\mathcal{S}})$.
- **Turn-based structures:** For any $h = 2t, t \in [H]$, we know that the private information are nested, due to the Assumption III.4 and construction of $\mathcal{D}'_{\mathcal{L}}$. We can apply the method of **Nested private information** part to computing the $\gamma_{1:h,h}^*$ with complexity poly$(\overline{\mathcal{P}}_h, \overline{\mathcal{A}}_h, \overline{\mathcal{S}})$.
  For any $h = 2t, t \in [H], \gamma_{ct(h),h} \in \Gamma_{ct(h)}, \gamma_{-ct(h),h}, \gamma'_{-ct(h),h} \in \Gamma_{-ct(h),h}$, where $ct(h)$ is the controller, it holds for any $\widehat{c}_h$ that

$$\begin{aligned}
&Q_h^{*,\mathcal{M}}(\widehat{c}_h, \gamma_{ct(h),h}, \gamma_{-ct(h),h}) \\
&= \sum_{\overline{s}_h, \overline{p}_h, \overline{s}_{h+1}, \overline{o}_{h+1}} \mathbb{P}_h^{\mathcal{M},c}(\overline{s}_h, \overline{p}_h \,|\, \widehat{c}_h) \overline{\mathbb{T}}_h(\overline{s}_{h+1} \,|\, \overline{s}_h, \gamma_{ct(h),h}(\overline{p}_{ct(h),h}) \gamma_{-ct(h),h}(\overline{p}_{-ct(h),h})) \\
&\overline{\mathbb{O}}_{h+1}(\overline{o}_{h+1} \,|\, \overline{s}_{h+1}) [\overline{\mathcal{R}}_h(\overline{s}_h, \gamma_{ct(h),h}(\overline{p}_{ct(h),h})) + V_{h+1}^{*,\mathcal{M}}(\widehat{c}_{h+1})] \\
&= \sum_{\overline{s}_h, \overline{p}_h, \overline{s}_{h+1}, \overline{o}_{h+1}} \mathbb{P}_h^{\mathcal{M},c}(\overline{s}_h, \overline{p}_h \,|\, \widehat{c}_h) \overline{\mathbb{T}}_h(\overline{s}_{h+1} \,|\, \overline{s}_h, \gamma_{ct(h),h}(\overline{p}_{ct(h),h}) \overline{\mathbb{O}}_{h+1}(\overline{o}_{h+1} \,|\, \overline{s}_{h+1}) [\overline{\mathcal{R}}_h(\overline{s}_h, \gamma_{ct(h),h}(\overline{p}_{ct(h),h})) + V_{h+1}^{*,\mathcal{M}}(\widehat{c}_{h+1})],
\end{aligned}$$

  where the last step is due to the fact that $\widehat{c}_{h+1} = \widehat{\phi}_{h+1}(\widehat{c}_h, \overline{z}_{h+1})$. And $\overline{z}_{h+1} = z_{\frac{h}{2}+1}^b = \chi_{\frac{h}{2}+1}(\overline{p}_h, \overline{a}_{ct(h),h}, \overline{o}_{h+1})$. Therefore, right-hand side does not depend on $\gamma_{-ct(h),h}$.
  In conclusion, Eq. (.30) with complexity poly$(\overline{\mathcal{S}}, \overline{\mathcal{P}}_{ct(h)}, \overline{\mathcal{A}}_{ct(h)})$.

- **Factorized structures:** For any $h \in [\overline{H}], t \in [H]$, for any $\widehat{c}_h \in \widehat{\mathcal{C}}_h, \gamma_h \in \Gamma_h$ we use backward induction to prove that, there exist n functions $\{F_{i,h}\}_{i\in[n]}$ such that

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

It holds for $h = \overline{H}+1$ obviously. For any $h \le \overline{H}$, it holds that

$$
\begin{aligned}
&Q_h^{*,\mathcal{M}}(\widehat{c}_h, \gamma_h) \\
&= \sum_{\overline{s}_h, \overline{p}_h, \overline{s}_{h+1}, \overline{o}_{h+1}} \mathbb{P}_h^{\mathcal{M},c}(\overline{s}_h, \overline{p}_h \mid \widehat{c}_h) \overline{\mathbb{T}}_h(\overline{s}_{h+1} \mid \overline{s}_h, \gamma_h(\overline{p}_h)) \overline{\mathbb{O}}_{h+1}(\overline{o}_{h+1} \mid \overline{s}_{h+1}) \\
&\qquad [\sum_{i=1}^{n} \overline{\mathcal{R}}_{i,h}(\overline{s}_{i,h}, \gamma_{i,h}(\overline{p}_{i,h}) + F_{i,h+1}(\widehat{c}_{i,h+1}, \widehat{g}^*_{i,h+1}(\widehat{c}_{i,h+1}))] \\
&= \sum_{i=1}^{n} \sum_{\overline{s}_{i,h}, \overline{p}_{i,h}, \overline{s}_{i,h+1}, \overline{o}_{i,h+1}} \mathbb{P}_{i,h}^{\mathcal{M},c}(\overline{s}_{i,h}, \overline{p}_{i,h} \mid \widehat{c}_{i,h}) \overline{\mathbb{T}}_h(\overline{s}_{i,h+1} \mid \overline{s}_{i,h}, \gamma_{i,h}(\overline{p}_{i,h})) \\
&\qquad \overline{\mathbb{O}}_{i,h+1}(\overline{o}_{i,h+1} \mid \overline{s}_{i,h+1}) [\overline{\mathcal{R}}_{i,h}(\overline{s}_{i,h}, \gamma_{i,h}(\overline{p}_{i,h}) + F_{i,h+1}(\widehat{c}_{i,h+1}, \widehat{g}^*_{i,h+1}(\widehat{c}_{i,h+1}))] \\
&=: \sum_{i=1}^{n} F_{i,h}(\widehat{c}_{i,h}, \gamma_{i,h}).
\end{aligned}
$$

Then, by induction, we know that it holds for any $h \in [\overline{H}]$. We can define $\widehat{g}^*_{i,h}(\widehat{c}_h) \in \operatorname{argmax}_{\gamma_{i,h} \in \Gamma_{i,h}} F_{i,h+1}(\widehat{c}_{i,h+1}, \gamma_{i,h})$, and thus solve Eq.(.30) with complexity $\sum_{i=1}^{n} \operatorname{poly}(\overline{\mathcal{S}}_i, \overline{\mathcal{A}}_{i,h}, \overline{\mathcal{P}}_{i,h})$.

This completes the proof. □

H. Examples in the Venn Diagram Fig. 2b

Here, we show some examples of the areas ①-⑤ in the Venn diagram in Fig. 2b.

- **①: Multi-agent MDP [35] with historical states.** The Dec-POMDPs satisfying that for any $h \in [H], i \in [n], \mathcal{O}_{i,h} = \mathcal{S}, \mathbb{O}_{i,h}(s \mid s) = 1, c_h = s_{1:h}, p_h = \emptyset$ lie in the area ①.
- **②: Uncontrolled state process without any historical information.** The Dec-POMDPs satisfying that for any $h \in [H], i \in [n], s_h, a_h, a'_h, \mathbb{T}_h(\cdot \mid s_h, a_h) = \mathbb{T}_h(\cdot \mid s_h, a'_h), c_h = \emptyset, p_{i,h} = \{o_{i,h}\}$ lie in the area ②.
- **③: Dec-POMDPs with sQC information structure and perfect recall, and satisfying Assumptions III.5 and III.7.** One-step delayed information sharing (Example 1 in A) lies in this area.
- **④: State controlled by one controller with no sharing and only observability of controller.** We consider a Dec-POMDP $\mathcal{D}$. The state dynamics are controller by only one agent (, for convenience, agent 1), and only agent 1 has observability, i.e. $\mathbb{T}_h(\cdot \mid s_h, a_{1,h}, a_{-1,h}) = \mathbb{T}_h(\cdot \mid s_h, a_{1,h}, a'_{-1,h})$ for all $s_h, a_{1,h}, a_{-1,h}, a'_{-1,h}$, and $\mathcal{O}_{-1,h} = \emptyset$. There is no information sharing, i.e. $c_h = \emptyset, p_{1,h} = \{o_{1:h}, a_{1:h-1}\}, p_{j,h} = \{a_{j,1:h-1}\}, \forall j \neq 1$. Then $\forall j \neq 1, h_1 < h_2 \in [H]$, agent $(1, h_1)$ does not influence (j, h_2), since $\tau_{j,h_2} = \{a_{j,1:h_2-1}\}$ is not influenced by agent $(1, h_1)$. Therefore, $\mathcal{D}$ is sQC and has perfect recall, $\mathcal{D}$ is not SI (underlying state s_h influenced by $g_{1,1:h-1}$). This is because $\mathcal{D}$ does not satisfy Assumption III.7. Then $\mathcal{D}$ lies in the area ④.
- **⑤: One-step delayed observation sharing and two-step delayed action sharing.** The Dec-POMDPs satisfying that for any $h \in [H], i \in [n], c_h = \{o_{1:h-1}, a_{1:h-2}\}, p_{i,h} = \{a_{i,h-1}, o_{i,h}\}$ lie in the area ⑤.

I. Additional Experimental Details and Results

a) Experimental setup: We conduct our experiments on two popular and modest-scale partially observable benchmarks, Dectiger [36] and Grid3x3 [37]. We train the agents in each LTC problem in the two environments with 20 different random seeds and different communication cost functions, and execute them in problems with horizons $[4, 6, 8, 10]$. To fit the setting of LTC in our paper. We regularize the reward between [0,1] and set the base information structure as one-step-delay. As for the communication cost function, we set $\mathcal{K}_h(Z_h^a) = \alpha |Z_h^a|$, and set $\alpha \in [0.01, 0.05, 0.1]$ for the purpose of ablation study. Also, we study 2 baselines under the same environment with information structure of one-step delay and fully-sharing, respectively. The one-step-delay baseline can be regarded as an LTC problem with extremely high communication cost, thus no additional sharing. On the other hand, the fully-sharing baseline is the LTC problem with no communication cost. Additionally, the results of different horizons and communications costs over 20 random seeds are shown in Tables I and II.

J. Additional Figures

We provide a few figures to better illustrate the paradigms and algorithmic ideas of this paper. Fig. 5 and Fig. 6 illustrate the paradigm and the timeline of the LTC problems considered in this paper.

Horizon/Cost	No Sharing	Cost=0.1	Cost=0.05	Cost=0.01	Fully Sharing
H=4 w/ cost	1.32±0.025	1.33±0.044	1.44±0.034	1.54±0.013	1.57±0.004
H=4 w/o cost	-	1.36±0.032	1.48±0.034	1.59±0.002	-
H=6 w/ cost	1.95±0.009	1.97±0.07	2.08±0.068	2.26±0.012	2.29±0.002
H=6 w/o cost	-	2.01±0.047	2.14±0.072	2.27±0.011	-
H=8 w/ cost	2.56±0.041	2.64±0.078	2.74±0.118	2.96±0.021	3.0±0.002
H=8 w/o cost	-	2.7±0.044	2.83±0.117	2.98±0.02	-
H=10 w/ cost	3.31±0.024	3.37±0.135	3.51±0.153	3.69±0.029	3.87±0.007
H=10 w/o cost	-	3.46±0.069	3.63±0.152	3.71±0.026	-

TABLE I: Experimental results for Dectiger.

Horizon/Cost	No Sharing	Cost=0.1	Cost=0.05	Cost=0.01	Fully Sharing
H=4 w/ cost	0.14±0.003	0.14±0.019	0.15±0.002	0.26±0.028	-0.48±0.023
H=4 w/o cost	-	0.14±0.019	0.21±0.007	0.33±0.023	-
H=6 w/ cost	0.33±0.02	0.32±0.025	0.4±0.009	0.48±0.059	-0.38±0.075
H=6 w/o cost	-	0.32±0.025	0.54±0.02	0.62±0.075	-
H=8 w/ cost	0.52±0.084	0.52±0.051	0.58±0.072	0.67±0.031	-0.4±0.022
H=8 w/o cost	-	0.52±0.051	0.72±0.035	0.82±0.074	-
H=10 w/ cost	0.73±0.02	0.73±0.037	0.9±0.169	1.03±0.019	-0.15±0.188
H=10 w/o cost	-	0.73±0.037	1.08±0.14	1.25±0.062	-

TABLE II: Experimental results for Grid3x3.

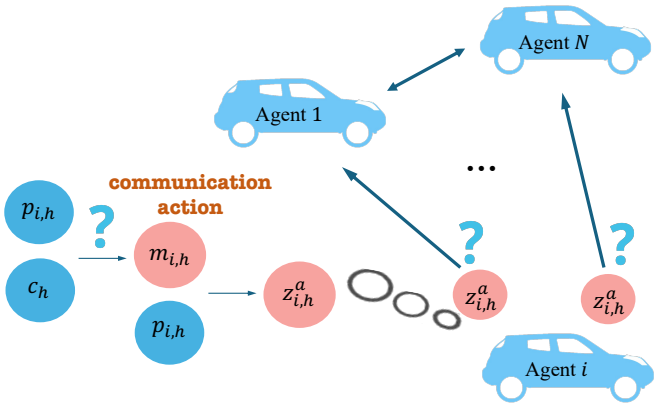


Fig. 5: Illustrating the paradigm of the Learning-to-Communicate problem considered in this paper.

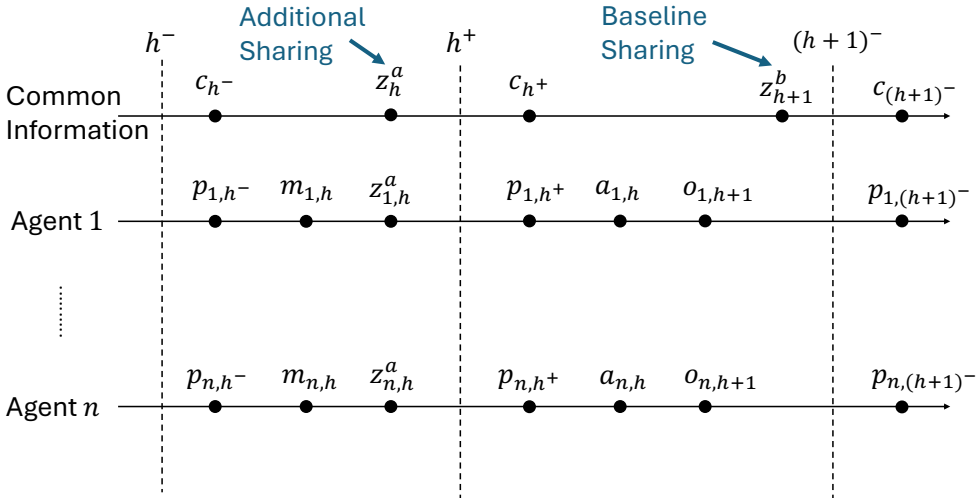


Fig. 6: Timeline of the information sharing and evolution protocols in the Learning-to-Communicate problem considered in this paper.