Principled Learning-to-Communicate in Cooperative MARL: An Information-Structure Perspective

Xiangyu Liu[†] **Haoyi You**[†] **Kaiqing Zhang**[†]
University of Maryland, College Park
`{xyliu999, yuriiyou, kaiqing}@umd.edu`

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 (CIB) framework, and classify LTCs 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 examples of QC LTCs satisfying the above conditions can be solved without computationally intractable oracles. Along the way, we also establish some relationships between (strictly) QC IS and the condition of strategy-independent CIB beliefs (SI-CIB), as well as solving general Dec-POMDPs without computationally intractable oracles beyond those with the SI-CIB condition, which may be of independent interest.

1 Introduction

Learning-to-communicate (LTC) 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 a *control* strategy that minimizes 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], inspiring us to examine LTCs from such a principled and rigorous perspective. Most of these studies, however, did not explore the *non-asymptotic* computational and/or sample complexity guarantees when the system knowledge is not (fully) known. A few recent studies [7, 8] started to explore the settings with general discrete (nonlinear) spaces, with special communication protocols and state transition dynamics.

More broadly, (the design of) communication strategy dictates the *information structure* (IS) of the control system, which characterizes *who knows what and when* [9]. IS and its impact on the *optimization tractability*, especially for linear systems, have been extensively studied in decentralized control, see [10, 11] 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.

[†]The authors are ordered alphabetically.

2nd Workshop on Aligning Reinforcement Learning Experimentalists and Theorists (ARLET 2025).

Specifically, we formalize LTCs in the general framework of decentralized partially observable Markov decision processes (Dec-POMDPs) [12], as in the empirical work [1, 2, 3]. To achieve finite-time and sample guarantees, we resort to the recent development in [13] on partially observable MARL, based on the common-information-based (CIB) framework [14, 15] 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 [14, 15, 13], allowing the sharing of *historical* information and communication costs; (ii) We classify LTCs through the lens of information structure, according to the ISs before additional information sharing. We then show that LTCs with *non-classical* [10] baseline IS is computationally intractable; (iii) Given the hardness, we 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) [15, 13], a condition previously shown to be critical for taming computational intractability [13]; (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, as well as solving general Dec-POMDPs beyond the SI-CIB ones, without computationally intractable oracles, and thus advancing the results in [13].

1.1 Related Work

Communication-control joint optimization. The joint design of control and communication strategies has been studied in the control literature [16, 6, 7, 8]. 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 [16, 6], or allowing to share only instantaneous observations [7, 8].

Information sharing and information structures. Information structure has been extensively studied to characterize *who knows what and when* in decentralized control [10, 11]. Our paper aims to formally understand LTC through the lens of information structures. The common-information-based approaches to formalize *information sharing* in [14, 15] 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, 12]. Recently, [20, 21] developed polynomial-sample complexity algorithms for partially observable stochastic games, but with computationally intractable oracles; [13] 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.

2 Preliminaries

2.1 Learning-to-Communicate (with Communication Cost)

For $n > 1$ agents, a (cooperative) *Learning-to-Communicate* problem is depicted 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. Other components are specified as follows.

2.1.1 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 for all the agents at timestep h. We denote by $\mathbb{T} = \{\mathbb{T}_h\}_{h\in[H]}$ the collection of state transition kernels, where $s_{h+1} \sim \mathbb{T}_h(\cdot \mid s_h, a_h) \in \Delta(\mathcal{S})$. We use $\mu_1 \in \Delta(\mathcal{S})$ to denote the initial state distribution. We denote 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 any state $s_h \in \mathcal{S}$. We denote by $\mathbb{O}_{i,h}(\cdot \mid s_h)$ the emission for agent i, the marginal distribution of $o_{i,h}$ given $\mathbb{O}_h(\cdot \mid s_h)$, at any $s_h \in \mathcal{S}$. At each timestep h, agents will receive a common reward $r_h = \mathcal{R}_h(s_h, a_h)$, where $\mathcal{R}_h : \mathcal{S} \times \mathcal{A}_h \to [0, 1]$ is the reward function shared by the agents.

2.1.2 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 z_h may contain two parts, the *baseline-sharing* part $z_h^b \in \mathcal{Z}_h^b$ that comes from some existing sharing protocol among agents, and the *additional-sharing* part $z_{i,h}^a \in \mathcal{Z}_{i,h}^a$ for each agent i that comes from explicit communication *to be decided/learned*, with the joint additional-sharing information $z_h^a := \cup_{i=1}^n z_{i,h}^a$. 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.

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. Hence, the *private information* of agent i at time h *before* and *after* additional sharing is defined accordingly as $p_{i,h^-} \subseteq \{o_{i,1:h}, a_{i,1:h-1}\} \backslash c_{h^-}, p_{i,h^+} \subseteq \{o_{i,1:h}, a_{i,1:h-1}\} \backslash c_{h^+}$. We denote by $p_{h^-} := (p_{1,h^-}, \cdots, p_{n,h^-})$ and $p_{h^+} := (p_{1,h^+}, \cdots, p_{n,h^+})$ the joint private information *before* and *after* additional sharing, respectively. 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 $m_{i,h} \in \mathcal{M}_{i,h}$ to denote the *communication action* of agent i at timestep h, determining what information $z_{i,h}^a$ she will share, through the way to be specified later. We denote 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. We use $\mathcal{K}_h : \mathcal{Z}_h^a \to [0, 1]$ to denote the *communication cost* function, and $\kappa_h = \mathcal{K}_h(z_h^a)$ to denote the communication cost at timestep h, due to additional sharing.

The evolution of information is formalized as follows.

Assumption 2.1 (*Information evolution*)**.**

(a) (Baseline sharing). For each $h \in [H-1]$, $z_{h+1}^b = \chi_{h+1}(p_{h^+}, a_h, o_{h+1})$ for some fixed transformation χ_{h+1}; for each agent $i \in [n]$, $p_{i,(h+1)^-} = \xi_{i,h+1}(p_{i,h^+}, a_{i,h}, o_{i,h+1})$ for some fixed transformation $\xi_{i,h+1}$, and the joint private information thus evolves as $p_{(h+1)^-} = \xi_{h+1}(p_{h^+}, a_h, o_{h+1})$ for some fixed transformation ξ_{h+1};

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

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

Note that the *fixed transformations* above (e.g., the χ_h and $\xi_{i,h}$) are not affected by the *realized values* of the random variables, but dictate some *pre-defined* transformation of the input random variables. See [14, 15] and §B in [13] for common examples of baseline sharing that admit such fixed transformations, and examples in §A in [22] on how they are extended to the LTC setting. Condition (c) above assumes that the agent has full memory of the information she has in the past and at present. We emphasize that this is closely related, but different from the common notion of *perfect recall* [23], where the agent has to also recall *all her past actions*. Condition (c), in contrast, relaxes the memorization of the actions, but includes the *instantaneous observation* $o_{i,h}$, as a basic requirement for *closed-loop* decision-making/control. This condition is satisfied by the models and examples in [10, 14, 15, 13], and see also §A in [22] for more examples that satisfy this assumption.

Meanwhile, for both the baseline and additional sharing protocols, we follow the model in the series of studies on partial history/information sharing [14, 15, 13, 7, 8] that, if an agent shares, she will share the information with *all other* agents. We make it formal below using the verbiage with σ-algebra, in order to be compatible with the literature on information structures [24, 10] to be discussed later.

Assumption 2.2. $\forall i_1, i_2 \in [n], h_1, h_2 \in [H], i_1 \neq i_2, h_1 < h_2$, 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 2.1-2.2 will be made throughout the paper.

2.1.3 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 *ϵ-team optimum* for $\mathcal{L}$ as follows.

Definition 2.3 (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

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

If $\epsilon = 0$, we call $(g_{1:H}^a, g_{1:H}^m)$ a team-optimal strategy of $\mathcal{L}$.

2.2 Information Structures of LTC

In decentralized stochastic control, the notion of information structure [24, 10] 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), and will be referred to as the *Dec-POMDP induced by* $\mathcal{L}$ (see a formal definition in [22, §E] for completeness).

In §2.1, we introduced LTC in the *state-space model*. Information structure, meanwhile, is usually more conveniently discussed within the equivalent framework of the *intrinsic model* [24]. In an intrinsic model, each agent only *acts once* throughout the problem 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 of LTCs in [22, §F].

(Strictly) quasi-classical ISs are important subclasses of ISs, which have been extensively studied in stochastic control [24, 25, 11] (see the instantiation for Dec-POMDPs in [22, §F]). We extend such a categorization to LTC problems with different ISs (of the baseline sharing) as follows.

Definition 2.4 ((Strictly) quasi-classical LTC)**.** We call an LTC $\mathcal{L}$ *(strictly) quasi-classical* if the *Dec-POMDP induced by* $\mathcal{L}$ *(denoted by* $\overline{\mathcal{D}}_{\mathcal{L}}$*),* i.e., the LTC problem without additional sharing, 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.

An LTC $\mathcal{L}$ that is not QC will thus be referred to as a *non-classical* LTC. See [22, §A] for the examples of QC/sQC LTC. Note that the categorization above is based on the ISs *before* additional sharing, an inherent property of the problem.

3 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` [12]. 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 are all deferred to [22, §B].

Recently, [26] showed that *observable* POMDPs, 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 generalized to Dec-POMDPs with information sharing in [13]. As solving LTCs is at least as hard as solving the Dec-POMDPs considered in [13], we first also make such an observability assumption on the *joint* emission function as in [13], to avoid computationally intractable oracles.

Assumption 3.1 (γ-observability [27, 26, 13])**.** 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 3.1 is not enough when it comes to LTC, if the baseline sharing IS is not favorable, in particular, *non-classical* [10]. The hardness persists even under a few additional assumptions to be introduced later that will make LTCs more tractable.

Lemma 3.2 (Non-classical LTCs are hard)**.** For non-classical LTCs under Assumptions 3.1, 3.4, 3.5, and 3.7, 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 3.1), even along with several other assumptions.

By Lemma 3.2, we will hence focus on the *quasi-classical* LTCs hereafter. Indeed, QC IS is also known to be critical for efficiently solving *linear* decentralized control [28, 29]. 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 §2.1.3 allows the dependence on agents' *private information*, which introduces incentives for *signaling* [10] and can also cause computational hardness, as shown next.

Lemma 3.3 (QC LTCs with full-history-dependent communication strategies are hard)**.** For QC LTCs under Assumption 3.1, together with Assumptions 3.5 and 3.7, 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 3.3 originates from the fact that when depending on the private/local information, determining the communication action can be made a *Team Decision Problem* [30], which is known to be hard. This will be the case even when the instantaneous observations are relatively observable (see Assumptions 3.1-3.7).

To avoid this hardness, we thus focus on communication strategies that only condition on the *common information*. Note that, this assumption does not lose the generality in the sense that the private information p_{i,h^-} *can still* be shared. It only means the communication action is not determined by 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 2.1), depending on p_{i,h^-}.

Assumption 3.4 (Common-information-based communication strategy)**.** The communication strategies take *common information* as input, with the following form:

$$\forall i \in [n], h \in [H], \quad g^m_{i,h} : \mathcal{C}_{h^-} \to \mathcal{M}_{i,h}. \tag{3.1}$$

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 the QC condition), while she starts to influence others by *sharing* her (*useless*) *control* actions, making 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.

Assumption 3.5 (Control-useless action is not used)**.** For each $i \in [n], h \in [H]$, if agent i's action $a_{i,h}$ does not influence the state s_{h+1}, namely, $\forall s_h \in \mathcal{S}, a_h \in \mathcal{A}_h, a'_{i,h} \in \mathcal{A}_{i,h}, a'_{i,h} \neq a_{i,h}, \mathbb{T}_h(\cdot \mid s_h, a_h) = \mathbb{T}_h(\cdot \mid s_h, (a'_{i,h}, a_{-i,h}))$. Then, $\forall h' > h, a_{i,h} \notin \tau_{h'-}$ and $a_{i,h} \notin \tau_{h'+}$.

Lemma 3.6 (QC LTCs without Assumption 3.5 are hard)**.** For QC LTCs under Assumptions 3.1, 3.4, and 3.7, finding a team-optimum is still `NP-hard`.

Note that Assumption 3.5 was *implicitly* made in the literature [15, 13] when there are *uncontrolled* state dynamics.

Assumption 3.7 (Other agents' emissions are non-degenerate)**.** For any $h \in [H], i \in [n]$, $\mathbb{O}_{-i,h}$ satisfies that $\forall b_1, b_2 \in \Delta(\mathcal{S}), b_1 \neq b_2,\ \mathbb{O}_{-i,h}^{\top} b_1 \neq \mathbb{O}_{-i,h}^{\top} b_2$.

Lemma 3.8 (QC LTCs without Assumption 3.7 are hard)**.** For QC LTCs under Assumptions 3.1, 3.4, and 3.5, 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 LTC in general. More importantly, as will be shown later, as another justification, LTCs under Assumptions 3.4, 3.5, and 3.7 can indeed *preserve* the QC/sQC information structure *after* additional sharing, making it possible for the overall LTC to be computationally tractable. Examples that satisfy these assumptions can also be found in [22, §A].

4 Solving LTC Problems Provably

We now study how to solve LTC provably, via either *planning* (with model knowledge) or *learning* (without model knowledge). The pipeline of our solution is shown in Fig. 1, and proofs of the results can be found in [22, §C].

4.1 An Equivalent Dec-POMDP $\mathcal{D}_{\mathcal{L}}$

Given any LTC $\mathcal{L}$, we can define a Dec-POMDP $\mathcal{D}_{\mathcal{L}}$ characterized by $\langle \widetilde{H}, \widetilde{\mathcal{S}}, \{\widetilde{\mathcal{A}}_{i,h}\}_{i\in[n],h\in[\widetilde{H}]}, \{\widetilde{\mathcal{O}}_{i,h}\}_{i\in[n],h\in[\widetilde{H}]}, \{\widetilde{\mathbb{T}}_h\}_{h\in[\widetilde{H}]}, \widetilde{\mathbb{O}}, \widetilde{\mu}_1, \{\widetilde{\mathcal{R}}_h\}_{h\in[\widetilde{H}]}\rangle$, such that these two are equivalent (under the assumptions in §3)

$$\begin{aligned}
&\widetilde{H} = 2H,\ \ \widetilde{\mathcal{S}} = \mathcal{S},\ \ \widetilde{s}_{2h-1} = \widetilde{s}_{2h} = s_h,\ \ \widetilde{\mathcal{A}}_{i,2h-1} = \mathcal{M}_{i,h},\ \ \widetilde{\mathcal{A}}_{i,2h} = \mathcal{A}_{i,h},\ \ \widetilde{\mathcal{O}}_{i,2h-1} = \mathcal{O}_{i,h},\\
&\widetilde{\mathcal{O}}_{i,2h} = \{\emptyset\},\ \ \widetilde{\mu}_1 = \mu_1,\ \ \widetilde{\mathbb{O}}_{2h-1} = \mathbb{O}_h,\ \ \widetilde{\mathbb{T}}_{2h-1}(\widetilde{s}_{2h} \mid \widetilde{s}_{2h-1}, \widetilde{a}_{2h-1}) = \mathbb{1}[\widetilde{s}_{2h} = \widetilde{s}_{2h-1}],\\
&\widetilde{\mathbb{T}}_{2h}(\widetilde{s}_{2h+1} \mid \widetilde{s}_{2h}, \widetilde{a}_{2h}) = \mathbb{T}_h(\widetilde{s}_{2h+1} \mid \widetilde{s}_{2h}, \widetilde{a}_{2h}),\ \ \widetilde{\mathcal{R}}_{2h-1} = -\mathcal{K}_h,\ \ \widetilde{\mathcal{R}}_{2h} = \mathcal{R}_h,\\
&\widetilde{p}_{i,2h-1} = \emptyset,\ \ \widetilde{p}_{i,2h} = p_{i,h^+},\ \ \widetilde{c}_{2h-1} = c_{h^-},\ \ \widetilde{c}_{2h} = c_{h^+},\ \ \widetilde{z}_{2h-1} = z_h^b,\ \ \widetilde{z}_{2h} = z_h^a,
\end{aligned} \tag{4.1}$$

Note that, at the odd timestep $2h-1$, we have $\widetilde{\tau}_{i,2h-1} := \widetilde{p}_{i,2h-1} \cup \widetilde{c}_{2h-1} = c_{h^-}$ under Assumption 3.4, 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 in $\mathcal{D}_{\mathcal{L}}$. Similarly, the objective of $\mathcal{D}_{\mathcal{L}}$ is defined as $J_{\mathcal{D}_{\mathcal{L}}}(\widetilde{g}_{1:\widetilde{H}}) = \mathbb{E}_{\mathcal{D}_{\mathcal{L}}}[\sum_{h=1}^{\widetilde{H}} \widetilde{r}_h \mid \widetilde{g}_{1:\widetilde{H}}]$. 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 [7, 8]. 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. One can verify that the $\mathcal{L}$ and $\mathcal{D}_{\mathcal{L}}$ are equivalent in terms of solution strategies, and $\mathcal{D}_{\mathcal{L}}$ inherits the QC IS from $\mathcal{L}$ (see [22, §IV.A]).

4.2 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 [28, 29]. With discrete spaces, the recent result [13] established a concrete *quai-polynomial*-time complexity for planning, under the *strategy* *independence* assumption [15] on the common-information-based beliefs [14, 15]. This SI-CIB assumption was shown critical for *computational tractability* [13] – 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 computation tractability.

To this end, one can first *expand* the QC $\mathcal{D}_{\mathcal{L}}$ to a *strictly* QC problem $\mathcal{D}_{\mathcal{L}}^{\dagger}$ 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. One can show that this *expansion* does not change the optimal value, and the (approximate) optimal strategy of $\mathcal{D}_{\mathcal{L}}^{\dagger}$ can be reduced to that of $\mathcal{D}_{\mathcal{L}}$ efficiently. More importantly, a benefit of having a *strictly* QC $\mathcal{D}_{\mathcal{L}}^{\dagger}$ is that, it has SI-CIBs under the assumptions in §3, making it possible to be solved without computationally intractable oracles as in [13]. See more details on this expansion in [22, §IV.B].

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

Despite having SI-CIBs, $\mathcal{D}_{\mathcal{L}}^{\dagger}$ is still not eligible for applying the results in [13]: the information evolution rules of $\mathcal{D}_{\mathcal{L}}^{\dagger}$ break those in [15, 13]. Specifically, due to Assumption 3.4, 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 [15, 13]. To address this issue, we propose to further *refine* $\mathcal{D}_{\mathcal{L}}^{\dagger}$ to obtain a Dec-POMDP $\mathcal{D}'_{\mathcal{L}}$, which satisfies the information evolution rules. The elements in $\mathcal{D}'_{\mathcal{L}}$ (denoted by the $\check{}$ notation) remain the same as those in $\mathcal{D}_{\mathcal{L}}^{\dagger}$, except that the private information at odd steps is now refined as $\overline{p}_{i,2t-1} = p_{i,t^-} \backslash \check{c}_{2t-1}$.

The new Dec-POMDP $\mathcal{D}'_{\mathcal{L}}$ is not equivalent to $\mathcal{D}_{\mathcal{L}}^{\dagger}$ 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 space, then solving $\mathcal{D}_{\mathcal{L}}^{\dagger}$ is equivalent to finding a *best-in-class* team-optimal strategy of $\mathcal{D}'_{\mathcal{L}}$ within the space $\overline{\mathcal{G}}_{1:\overline{H}}$, as shown below.

Theorem 4.1. Let $\mathcal{D}_{\mathcal{L}}^{\dagger}$ be an sQC Dec-POMDP generated from $\mathcal{L}$ after reformulation and strict expansion, and $\mathcal{D}'_{\mathcal{L}}$ be the refinement of $\mathcal{D}_{\mathcal{L}}^{\dagger}$ as above. Then, finding the optimal strategy in $\mathcal{D}_{\mathcal{L}}^{\dagger}$ 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 $h \in [\overline{H}]$:

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

with some functions $\{\overline{\chi}_{h+1}\}_{h\in[\overline{H}]}, \{\overline{\xi}_{i,h+1}\}_{i\in[n],h\in[\overline{H}]}$.Furthermore, if Assumptions 3.4, 3.5 and 3.7 hold, then $\mathcal{D}'_{\mathcal{L}}$ has SI-CIBs with respect to the strategy spaces $\overline{\mathcal{G}}_{1:\overline{H}}$, 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 from both $\overline{g}_{1:h-1}$ and $\overline{g}'_{1:h-1}$, it holds that $\mathbb{P}_h^{\mathcal{D}'_{\mathcal{L}}}(\overline{s}_h, \overline{p}_h \,|\, \overline{c}_h, \overline{g}_{1:h-1}) = \mathbb{P}_h^{\mathcal{D}'_{\mathcal{L}}}(\overline{s}_h, \overline{p}_h \,|\, \overline{c}_h, \overline{g}'_{1:h-1})$.

4.4 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 [13]. Given a Dec-POMDP $\mathcal{D}'_{\mathcal{L}}$ with SI-CIBs, [13] proposed to construct an *expected-approximate common information model* $\mathcal{M}$ through *finite memory* (as defined in [22, §C]), when $\mathcal{D}'_{\mathcal{L}}$ is γ-observable. However, the Dec-POMDP $\mathcal{D}'_{\mathcal{L}}$ obtained from LTC has two key differences from the general ones considered in [13]. 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 [13] 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 [13]. Hence, we can leverage the approaches in [13] to find an approximately optimal strategy $\overline{g}^*_{1:\overline{H}}$ by finding an optimal prescriptions $\gamma^*_{1:\overline{H}}$ under each possible $\widehat{c}_{1:\overline{H}}$ with backward induction from timesteps $h = \overline{H}$ to 1.

Note that in each step of the backward induction, a *Team Decision Problem* (TDP) [30] needs to be solved for each $\widehat{c}_h$, which is known to be `NP-hard` in general [30]:

$$\big(\widehat{g}^*_{1,h}(\cdot\,|\,\widehat{c}_h,\cdot),\cdots,\widehat{g}^*_{n,h}(\cdot\,|\,\widehat{c}_h,\cdot)\big) \leftarrow \operatorname*{argmax}_{\gamma_h} Q_h^{*,\mathcal{M}}(\widehat{c}_h,\gamma_h), \tag{4.2}$$

where the Q-value function and the prescriptions γ_h are defined in [22, §C]. Hence, to obtain overall computational tractability, we make the following *one-step* tractability assumption, as in [13].

Assumption 4.2 (One-step tractability of $\mathcal{M}$)**.** The one-step Team Decision Problems induced by $\mathcal{M}$ (i.e., Eq. (4.2)) 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 3.4, it suffices to only assume one-step tractability for the *control* (i.e., even) steps. Third, even without Assumption 4.2, 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 4.2). Finally, such an assumption is satisfied for several classes of Dec-POMDPs with information sharing, see [22, §G] for more examples. With this assumption, we can obtain a planning algorithm with quasi-polynomial time complexity (see [22, §C]).

4.5 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 (§4.1) does not change the system dynamics, but only maps the information to different random variables; the *expansion* step (§4.2) only requires agents to share more actions with each other, without changing the input and output of the environment; the *refinement* step (§4.3) only recovers the private information the agents had in the original $\mathcal{L}$. This way, we can utilize similar algorithmic ideas in [13] to develop provable learning algorithms for LTC problems. See [22, §C] for more details of the provable LTC algorithm adapted from [13]. The algorithm has the following finite-time and sample complexity guarantees.

Theorem 4.3. Given any QC LTC problem $\mathcal{L}$ satisfying Assumptions 3.1, 3.4, 3.5, and 3.7, we can construct an SI-CIB Dec-POMDP problem $\mathcal{D}'_{\mathcal{L}}$. Moreover, there exists an LTC algorithm (see [22, §C]) learning in $\mathcal{D}'_{\mathcal{L}}$, such that if the expected-approximate common information models $\widehat{\mathcal{M}}$ in the algorithm can be constructed and satisfy Assumption 4.2, 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 [22, §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.

5 Solving General QC Dec-POMDPs

In §4, 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 [13] that can only address *SI-CIB* Dec-POMDPs, a result of independent interest. Without much confusion given the context, we will adapt the notation of LTC to studying general Dec-POMDPs: we set $h^+ = h^- = h$ and void the additional sharing protocol. We extend the results in §4 to general QC Dec-POMDPs as follows.

Theorem 5.1. Consider a Dec-POMDP $\mathcal{D}$ under Assumptions 2.1 (c). If $\mathcal{D}$ is sQC and satisfies Assumptions 2.2, 3.5, and 3.7, 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 [23] here means that the agents will never forget their own past information *and actions* (as formally defined in [22, §D]). Note that Assumption 2.1 (c) is similar but different from perfect recall: it is implied by the latter with $o_{i,h} \in \tau_{i,h}$. Also, Assumptions 2.2, 3.5, and 3.7 were originally made for LTCs, and here we meant to impose them for Dec-POMDPs with $h^+ = h^- = h$. Finally, by

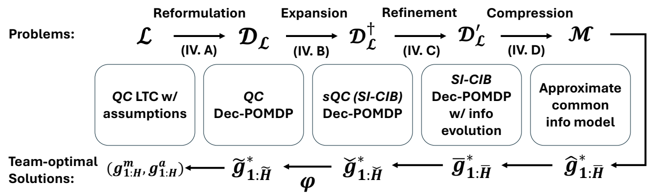
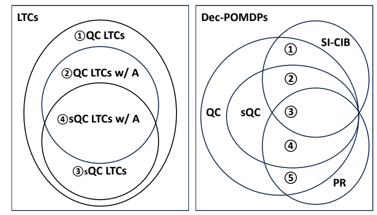


Figure 1: Illustrating the subroutines for solving the LTC problems.



(a) (b)

Figure 2: Venn diagrams

In Fig. 2, (a) Venn diagram of LTCs with different ISs: ① QC LTCs. ② QC LTCs satisfying Assumptions 3.4, 3.5, and 3.7. ③ sQC LTCs. ④ sQC LTCs satisfying Assumptions 3.4, 3.5, and 3.7, whose reformulated Dec-POMDPs have SI-CIB; (b) Venn diagram of general Dec-POMDPs with different ISs. PR denotes perfect recall. ③ denotes the Dec-POMDPs we mainly consider, e.g., the examples in [15, 13]. We also construct examples for other areas in [22, §H]

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 (as in $\mathcal{D}'_{\mathcal{L}}$). Given Theorem 5.1 and the results in §4, we illustrate the relationship between LTCs and Dec-POMDPs with different assumptions and ISs in Fig. 2, which may be of independent interest.

6 Experimental Results

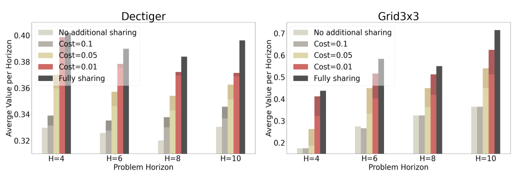


Figure 3: The average values achieved under different communication costs and horizons.

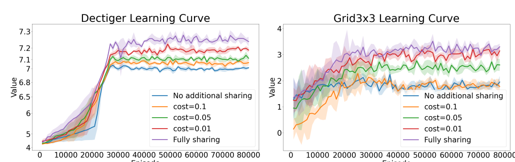


Figure 4: Learning curves with different communication costs.

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 in [22]. The attained average-values are presented in Fig. 3, where 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. 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.

7 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.

Acknowledgement

The authors acknowledge the support from Army Research Office (ARO) grant W911NF-24-1-0085, NSF CAREER Award 2443704, and AFOSR YIP Award FA 9550-25-1-0258.

References

[1] Jakob Foerster, Ioannis Alexandros Assael, Nando De Freitas, and Shimon Whiteson. Learning to communicate with deep multi-agent reinforcement learning. In *NeurIPS*, 2016.

[2] Sainbayar Sukhbaatar, Rob Fergus, et al. Learning multiagent communication with backpropagation. In *NeurIPS*, 2016.

[3] Jiechuan Jiang and Zongqing Lu. Learning attentional communication for multi-agent cooperation. In *NeurIPS*, 2018.

[4] Sekhar Tatikonda and Sanjoy Mitter. Control under communication constraints. *IEEE Trans. Autom. Control*, 49:1056–1068, 2004.

[5] Girish N Nair, Fabio Fagnani, Sandro Zampieri, and Robin J Evans. Feedback control under data rate constraints: An overview. *Proceed. of the IEEE*, 95:108–137, 2007.

[6] Serdar Yüksel. Jointly optimal LQG quantization and control policies for multi-dimensional systems. *IEEE Trans. Autom. Control*, 59:1612–1617, 2013.

[7] Sagar Sudhakara, Dhruva Kartik, Rahul Jain, and Ashutosh Nayyar. Optimal communication and control strategies in a multi-agent mdp problem. *arXiv preprint arXiv:2104.10923*, 2021.

[8] Dhruva Kartik, Sagar Sudhakara, Rahul Jain, and Ashutosh 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.

[9] Hans S Witsenhausen. Separation of estimation and control for discrete time systems. *Proceed. of the IEEE*, 59:1557–1566, 1971.

[10] Aditya Mahajan, Nuno C Martins, Michael C Rotkowitz, and Serdar Yüksel. Information structures in optimal decentralized control. In *IEEE Conf. on Dec. and Control*, 2012.

[11] Serdar Yüksel and Tamer Başar. *Stochastic Teams, Games, and Control under Information Constraints*. Springer Nature, 2023.

[12] Daniel S Bernstein, Robert Givan, Neil Immerman, and Shlomo Zilberstein. The complexity of decentralized control of markov decision processes. *Math. Oper. Res.*, 27:819–840, 2002.

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

[14] Ashutosh Nayyar, Aditya Mahajan, and Demosthenis Teneketzis. Decentralized stochastic control with partial history sharing: A common information approach. *IEEE Trans. Autom. Control*, 58(7):1644–1658, 2013.

[15] Ashutosh Nayyar, Abhishek Gupta, Cedric Langbort, and Tamer Başar. Common information based Markov perfect equilibria for stochastic games with asymmetric information: Finite games. *IEEE Trans. Autom. Control*, 59:555–570, 2013.

[16] Lin Xiao, Mikael Johansson, Haitham Hindi, Stephen Boyd, and Andrea Goldsmith. Joint optimization of wireless communication and networked control systems. *Switching and Learning Feedback Sys.*, pages 248–272, 2005.

[17] Christos H Papadimitriou and John N Tsitsiklis. The complexity of Markov decision processes. *Math. Oper. Res.*, 12:441–450, 1987.

[18] Christopher Lusena, Judy Goldsmith, and Martin Mundhenk. Nonapproximability results for partially observable Markov decision processes. *J. Artif. Intell. Res.*, pages 83–103, 2001.

[19] Chi Jin, Sham Kakade, Akshay Krishnamurthy, and Qinghua Liu. Sample-efficient reinforcement learning of undercomplete pomdps. In *NeurIPS*, 2020.

[20] Qinghua Liu, Csaba Szepesvári, and Chi Jin. Sample-efficient reinforcement learning of partially observable Markov games. In *NeurIPS*, 2022.

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

[22] Xiangyu Liu, Haoyi You, and Kaiqing Zhang. Principled learning-to-communicate with quasi-classical information structures. 2025. `https://kzhang66.github.io/assets/LTC_full.pdf`.

[23] Harold W. Kuhn. Extensive games and the problem of information. In *Contrib. Theory Games, Vol. II*. Princeton Univ. Press, 1953.

[24] Hans S Witsenhausen. The intrinsic model for discrete stochastic control: Some open problems. In *Control Theory, Numer. Methods Comput. Syst. Model., Int. Symp., Rocquencourt*, pages 322–335, 1975.

[25] Aditya Mahajan and Serdar Yüksel. Measure and cost dependent properties of information structures. In *Amer. Control Conf.*, pages 6397–6402, 2010.

[26] Noah Golowich, Ankur Moitra, and Dhruv Rohatgi. Planning and learning in partially observable systems via filter stability. In *Proc. 55th Annu. ACM Symp. Theory Comput.*, 2023.

[27] Eyal Even-Dar, Sham M Kakade, and Yishay Mansour. The value of observation for monitoring dynamic systems. In *IJCAI*, 2007.

[28] Yu-Chi Ho et al. Team decision theory and information structures in optimal control problems – part i. *IEEE Trans. Autom. Control*, 17:15–22, 1972.

[29] Andrew Lamperski and Laurent Lessard. Optimal decentralized state-feedback control with sparsity and delays. *Automatica*, pages 143–151, 2015.

[30] John Tsitsiklis and Michael Athans. On the complexity of decentralized decision making and detection problems. *IEEE Trans. Autom. Control*, 30:440–446, 1985.