Provable Partially Observable Reinforcement Learning with Privileged Information

Yang Cai[1], **Xiangyu Liu**[2], **Argyris Oikonomou**[1], **Kaiqing Zhang**[2]
[1]Yale University, [2]University of Maryland, College Park
yang.cai@yale.edu,xyliu@umd.edu
argyris.oikonomou@yale.edu, kaiqing@umd.edu

Abstract

Partial observability of the underlying states generally presents significant challenges for reinforcement learning (RL). In practice, certain *privileged information*, e.g., the access to states from simulators, has been exploited in training and achieved prominent empirical successes. To better understand the benefits of privileged information, we revisit and examine several simple and practically used paradigms in this setting, with both computation and sample efficiency analyses. Specifically, we first formalize the empirical paradigm of *expert distillation* (also known as *teacher-student* learning), demonstrating its pitfall in finding near-optimal policies. We then identify a condition of the partially observable environment, the deterministic filter condition, under which expert distillation achieves sample and computational complexities that are *both* polynomial. Furthermore, we investigate another successful empirical paradigm of *asymmetric actor-critic*, and focus on the more challenging setting of observable partially observable Markov decision processes. We develop a belief-weighted optimistic asymmetric actor-critic algorithm with polynomial sample and quasi-polynomial computational complexities, where one key component is a new provable oracle for learning belief states that preserve *filter stability* under a misspecified model, which may be of independent interest. Finally, we also investigate the provable efficiency of partially observable multi-agent RL (MARL) with privileged information. We develop algorithms with the feature of centralized-training-with-decentralized-execution, a popular framework in empirical MARL, with polynomial sample and (quasi-)polynomial computational complexity in both paradigms above. Compared with a few recent related theoretical studies, our focus is on understanding practically inspired algorithmic paradigms, without computationally intractable oracles.

1 Introduction

In most real-world applications of reinforcement learning (RL), e.g., perception-based robot learning [38, 2], dialogue systems [78], and clinical trials [68], only partial observations of the environment state are available for sequential decision-making. Such partial observability presents significant challenges for efficient decision-making, with known computational [57] and statistical [35, 31] barriers under the general model of partially observable Markov decision processes (POMDPs). The curse of partial observability becomes severer when *multiple* RL agents interact, where not only the environment state, but also other agents' information, are unobservable in decision-making [75, 71].

On the other hand, a flurry of empirical paradigms has made partially observable (multi-agent) RL promising in practice. One notable example is to exploit the *privileged information* that may be available (only) during training. The privileged information usually includes direct access to the underlying states, as well as access to other agents' observations/actions in multi-agent RL (MARL), due to the use of simulators and/or high-precision sensors for training. The latter is also

38th Workshop on Aligning Reinforcement Learning Experimentalists and Theorists (ARLET 2024).

known as the *centralized-training-with-decentralized-execution* (CTDE) framework, and has become prevalent in deep MARL [44, 61, 18]. These approaches can be mainly categorized into two types: i) privileged *policy* learning, where an expert/teacher policy is trained with privileged information, and then *distilled* into a student partially observable policy. This *expert distillation*, also known as *teacher-student learning*, approach has been the key to empirical successes in robotic locomotion [37, 50] and autonomous driving [11]; ii) privileged *value* learning, where a value function is trained conditioned on privileged information, and used to improve a partially observable policy. It is typically instantiated as the asymmetric actor-critic/policy optimization algorithm [59], and serves as the backbone of the high-profiled successes in robotic manipulation [38, 2] and MARL [44].

Despite the remarkable empirical successes, theoretical understandings of partially observable RL with privileged information have been rather limited, except for a few recent prominent advances in RL with hindsight observability [36, 26] (see Appendix B for a detailed discussion). However, most of these theoretically sound algorithms are different from those used in practice, requiring computationally intractable oracles to achieve provable sample efficiency. The soundness and efficiency of the aforementioned paradigms used in practice remain elusive. In this work, we examine both paradigms of expert distillation and asymmetric actor-critic, with foresight privileged information as in these empirical works. In contrast to [36, 26], which purely focused on sample efficiency, we aim to understand the benefits of privileged information by examining these practically inspired paradigms under several POMDP models, without computationally intractable oracles. We defer a detailed literature review to Appendix B and summarize our contribution as follows.

Contributions. We first formalize the empirical paradigm of expert distillation, and demonstrate its pitfall in distilling near-optimal policies even in observable POMDPs, a model class that admits provable partially observable RL without computationally intractable oracles [21]. We then identify a new condition for POMDPs, the *deterministic filter* condition, and establish sample and computational complexities that are *both* polynomial for expert distillation. The new condition is weaker and thus encompasses several known (statistically) tractable POMDP models (see Figure 1 for a figurative summary). Further, we revisit the asymmetric actor-critic paradigm and analyze its efficiency under the more challenging setting of observable POMDPs (where expert distillation fails). Observing the inefficiency of vanilla asymmetric actor-critic, and inspired by the empirical success in belief-state-learning, we develop a new optimistic version of asymmetric actor-critic, with polynomial-sample and quasi-polynomial-time complexities. Key to the results is a new belief-state learning oracle that preserves filter stability under a misspecified model, which may be of independent interest. Finally, we also investigate the provable efficiency of partially observable MARL with privileged information, by studying algorithms under the CTDE framework, with polynomial-sample and (quasi-)polynomial-time complexities in both paradigms above.

2 Preliminaries

2.1 Partially Observable RL (with Privileged Information)

Model. Formally, we define a POMDP by a tuple $\mathcal{P} = (H, \mathcal{S}, \mathcal{A}, \mathcal{O}, \mathbb{T}, \mathbb{O}, \mu_1, r)$, where H denotes the length of each episode, $\mathcal{S}$ is the state space with $|\mathcal{S}| = S$, $\mathcal{A}$ denotes the action space with $|\mathcal{A}| = A$. We use $\mathbb{T} = \{\mathbb{T}_h\}_{h\in[H]}$ to denote the collection of the transition matrices, so that $\mathbb{T}_h(\cdot \mid s, a) \in \Delta(\mathcal{S})$ gives the probability of the next state if action a is taken at state s and step h. In the following discussions, for any given a, we treat $\mathbb{T}_h(a) \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ as a matrix, where each row gives the probability for the next state. We use μ_1 to denote the distribution of the initial state s_1, and $\mathcal{O}$ to denote the observation space with $|\mathcal{O}| = O$. We use $\mathbb{O} = \{\mathbb{O}_h\}_{h\in[H+1]}$ to denote the collection of the joint emission matrices, so that $\mathbb{O}_h(\cdot \mid s) \in \Delta(\mathcal{O})$ gives the emission distribution over the joint observation space $\mathcal{O}$ at state s and step h. For notational convenience, we will at times adopt the matrix convention, where $\mathbb{O}_h$ is a matrix with rows $\mathbb{O}_h(\cdot \mid s_h)$. Finally, $r = \{r_h\}_{h\in[H]}$ is a collection of reward functions, so that $r_h(s_h, a_h) \in [0, 1]$ is the reward given the state s_h and action a_h at step h. We thus denote the trajectory with states as $\overline{\tau}_h = (s_{1:h}, o_{1:h}, a_{1:h-1})$, the trajectory without states as $\tau_h = (o_{1:h}, a_{1:h-1})$, and its space as $\mathcal{T}_h$. Finally, we use $\boldsymbol{b}_h(\tau_h)$ to denote the posterior distribution over the latent state at step h given history τ_h, which is known as the *belief state* (c.f. Appendix C.1 for a more detailed introduction).

Policy and value function. We define a stochastic policy at step h as:

$$\pi_h : \mathcal{O}^h \times \mathcal{A}^{h-1} \to \Delta(\mathcal{A}), \tag{2.1}$$

	Without PI	With PI (Ours)
Block MDP	Oracle-efficient [30] Computationally harder than SL [24]	Tabular: poly sample poly time FA: poly sample + classification oracle
k-decodable POMDP	Exponential in k sample and time [14]	
Det. POMDP	Without WSE: statistically hard [39] With WSE: poly sample+time [31]	
POSG with det. filter	N/A	poly sample poly time
Observable POMDP	quasi-poly sample + time [21] [43]	poly sample quasi-poly time
Observable POSG		

Table 1: Comparison of the theoretical guarantees with and without privileged information. PI: privileged information; SL: supervised learning; FA: function approximation; WSE: well-separated emission.

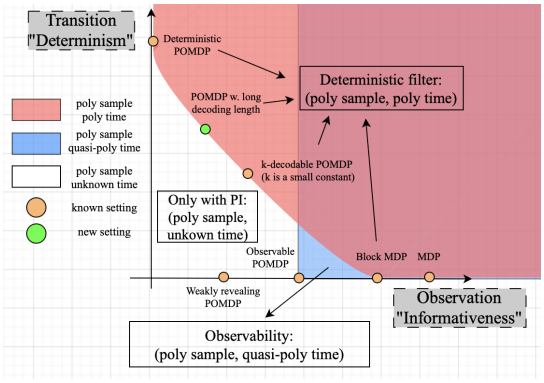

Figure 1: Landscape of POMDP models that partially observable RL with privileged information can/cannot address. The axes denote the "restrictiveness" of the assumptions, on the emission channels and transition dynamics, respectively.

where the agent bases on the entire history for decision-making. The corresponding policy class is denoted as Π_h. We further denote $\Pi = \times_{h\in[H]}\Pi_h$. Finally, we define $\Pi^{\text{gen}} := \{\pi_{1:H} \,|\, \pi_h : \mathcal{S}^h \times \mathcal{O}^h \times \mathcal{A}^{h-1} \to \Delta(\mathcal{A}) \text{ for } h \in [H]\}$ to be the most general policy space in partially observable RL with privileged state information, which can potentially depend on all historical states, observations, and actions. It can be seen that $\Pi \subseteq \Pi^{\text{gen}}$. We may also use policies that only receive a *finite memory* instead of the whole history as inputs. Formally, fix $L > 0$, we define the policy space Π^L to be the space of all possible policies $\{\pi_h\}_{h\in[H]}$ such that $\pi_h : \mathcal{Z}_h \to \Delta(\mathcal{A})$ with $\mathcal{Z}_h := \mathcal{O}^{\min\{L,h\}} \times \mathcal{A}^{\min\{L,h\}}$ for each $h \in [H]$. Finally, we define the space of policies that are only based on the state as $\Pi_{\mathcal{S}}$.

Given the POMDP model $\mathcal{P}$, we define the value function as $V_h^{\pi,\mathcal{P}}(y_h) := \mathbb{E}_\pi^{\mathcal{P}}[\sum_{t=h}^H r_t(s_t, a_t) \,|\, y_h]$, indicating the expected accumulated rewards from step h, where $y_h \subseteq (s_{1:h}, o_{1:h}, a_{1:h-1})$, and we abuse notation by treating as a set the sequence of states $s_{1:h}$, the sequence of observations $o_{1:h}$, and the sequence of actions $a_{1:h-1}$ up to time h, which are the available information at step h. We say y_h is *reachable* if there exists some policy $\pi \in \Pi^{\text{gen}}$ such that $\mathbb{P}^{\pi,\mathcal{P}}(y_h) > 0$. For $h = 1$, we adopt the simplified notation $v^{\mathcal{P}}(\pi) = \mathbb{E}_\pi^{\mathcal{P}}[\sum_{h=1}^H r_h(s_h, a_h)]$. Meanwhile, we also define $Q_h^{\pi,\mathcal{P}}(y_h, a_h) := \mathbb{E}_\pi^{\mathcal{P}}[\sum_{t=h}^H r_t(s_t, a_t) \,|\, y_h, a_h]$. We denote $d_h^{\pi,\mathcal{P}}(s_h) = \mathbb{P}^{\pi,\mathcal{P}}(s_h)$ to be the occupancy measure on the state space. The goal of learning in POMDPs is to find the optimal policy that maximizes the expected accumulated reward. Formally, we define:

Definition 2.1 (ϵ-optimal policy)**.** Given $\epsilon > 0$, a policy $\pi^\star \in \Pi$ is ϵ-optimal, if $v^{\mathcal{P}}(\pi^\star) \geq \max_{\pi\in\Pi} v^{\mathcal{P}}(\pi) - \epsilon$.

Learning with privileged information. Common RL algorithms for POMDPs deal with the scenario where during both training and test time, the agent can only observe its historical observations and actions, while the states are not accessible. In other words, the agent can only utilize policies from Π to interact with the environment. In contrast, in settings with privileged information, e.g., training in simulators and/or using sensors with higher precision, the underlying state can be used in learning. Thus, the agent is allowed to utilize policies from the class Π^{gen} during training. Meanwhile, the objective is still to find the optimal history-dependent policy in the space of Π, since at test time, the agent cannot access the state information anymore.

2.2 Partially Observable MARL with Information Sharing

Partially observable stochastic games (POSGs) are a natural generalization of POMDPs with multiple agents of potentially independent interests. We define a POSG with n agents by a tuple $\mathcal{G} = (H, \mathcal{S}, \{\mathcal{A}_i\}_{i=1}^n, \{\mathcal{O}_i\}_{i=1}^n, \mathbb{T}, \mathbb{O}, \mu_1, \{r_i\}_{i=1}^n)$, where each agent i has its individual action space $\mathcal{A}_i$, observation space $\mathcal{O}_i$, and reward function r_i. Particularly, an episode of POSG executes as follows: at each step h and state s_h, an observation is drawn from $(o_{i,h})_{i\in[n]} \sim \mathbb{O}_h(\cdot \,|\, s_h)$, and each agent receives its own observation $o_{i,h}$, takes corresponding action $a_{i,h}$, obtain the reward $r_{i,h}(s_h, a_h)$, where $a_h := (a_{i,h})_{i\in[n]}$, and then the system transitions into the next state. Notably, each agent i may not only know its local information $(o_{i,1:h}, a_{i,1:h-1})$, but also information from some other agents.

Therefore, we denote the information available to each agent i at step h also as $\tau_{i,h} \subseteq (o_{1:h}, a_{1:h-1})$ and define the *common information* as $c_h = \cap_{i\in[n]}\tau_{i,h}$ and *private information* as $p_{i,h} = \tau_{i,h} \setminus c_h$. We denote the space for common information and private information as $\mathcal{C}_h$ and $\mathcal{P}_{i,h}$ for each agent i and step h. We refer more examples to Appendix C.2. Correspondingly, the policy each agent i deploys at test time takes the form of $\pi_{i,h} : \Omega_h \times \mathcal{C}_h \times \mathcal{P}_{i,h} \to \Delta(\mathcal{A}_i)$, where Ω_h is the space of random seeds. We denote the policy space for agent i as Π_i. If $\pi_{i,h}$ takes the state instead of $c_h, p_{i,h}$ as input, we denote its policy space as $\Pi_{\mathcal{S},i}$. Note that this model covers the several recent POSG models studied for partially observable MARL, e.g., [41, 23]. For example, if there is no shared information, then $c_h = \emptyset$, and if all history information is shared, then $p_{i,h} = \emptyset$. In privileged-information-based learning, the training algorithm may exploit not only the underlying state information, but also the observations and actions of other agents.

Solution concepts. Different from POMDPs, where we hope to learn an optimal policy, the solution concepts for POSGs are the equilibria, particularly Nash equilibrium (NE) for two-player zero-sum games (i.e., when $n = 2$ and $r_{1,h} + r_{2,h} = 0$), and correlated equilibrium (CE) or coarse correlated equilibrium (CCE) for general-sum games. We defer a formal definition of them to Appendix C.2.

2.3 Technical Assumptions for Computational Tractability

In order to circumvent the known computational hardness of POMDPs/POSGs, we here rely on some standard and well-known assumptions, which include γ-observability, Assumption C.8 [16, 22, 21], and strategy independence of belief for POSGs, Assumption C.9 [53, 27, 43], for which we defer the formal introductions to Appendix C.3.

3 Revisiting Empirical Paradigms of RL with Privileged Information

Most empirical paradigms of RL with privileged information can be categorized into two types: i) privileged *policy* learning, where the policy in training is conditioned on the privileged information, and the trained policy is then *distilled* to a policy that does not take state as input. This is usually referred to as either expert distillation [11, 55, 49] or teacher-student learning [37, 50, 66] in the literature; ii) privileged *value* learning, where the value function is conditioned on the privileged information, and is then used to directly output a policy that takes observation (history) as input. One prominent example of ii) is asymmetric-actor-critic [59, 2]. It is worth noting that asymmetric-actor-critic is also closely related to one of the most successful paradigms for multi-agent RL, centralized-training-with-decentralized-execution [44, 76, 17]. Here we formalize and revisit the potential pitfalls of these two paradigms, and further develop our theoretically sound algorithms.

3.1 Privileged Policy Learning: Expert Policy Distillation

The motivation behind expert policy distillation is that learning an optimal fully observable policy in MDP is a much easier and better-studied problem with a bunch of well-known efficient algorithms. The (expected) distillation objective can be formalized as follows:

$$\widehat{\pi}^\star \in \arg\min_{\pi\in\Pi} \ \mathbb{E}^{\mathcal{P}}_{\pi'}\left[\sum_{h=1}^{H} D_f(\pi^\star_h(\cdot \mid s_h) \mid \pi_h(\cdot \mid \tau_h))\right], \tag{3.1}$$

where π' is some given behavioral policy to collect exploratory trajectories, $\pi^\star \in \arg\max_{\pi\in\Pi_{\mathcal{S}}} v^{\mathcal{P}}(\pi)$ denotes the optimal fully observable policy, and D_f denotes the general f-divergence to measure the discrepancy between $\pi^\star$ and π.

Such a formulation looks promising since it essentially circumvents the challenging issue of exploration in partially observable environments, by directly mimicking an expert policy that can be obtained from any off-the-shelf MDP learning algorithms. However, we point out in the following proposition that even if the POMDP satisfies Assumption C.8, the distilled policy can still be strictly suboptimal even with infinite data, i.e., by solving the expected objective Equation (3.1) directly. We postpone the proof of Proposition 3.1 to Appendix E.

Proposition 3.1 (Pitfall of expert policy distillation)**.** For any $\epsilon, \gamma \in (0, 1)$, there exists an γ-observable POMDP $\mathcal{P}^\epsilon$ with $H = 1$, $S = O = A = 2$ such that for any behavioral policy π' and choice of D_f in Equation (3.1), it holds that $v^{\mathcal{P}^\epsilon}(\widehat{\pi}^\star) \le \max_{\pi\in\Pi} v^{\mathcal{P}^\epsilon}(\pi) - \frac{(1-\epsilon)(1-\gamma)}{4}$.

The key reason why Equation (3.1) fails is that the underlying state can remain highly uncertain even given the history information. Thus, the distilled partially observable policy may never be able to mimic the expert very well.

To see how we may rule out such problems, we notice that if $\gamma = 1$,[1], implying that the observation can decode the latent state exactly, the bound in Proposition 3.1 becomes vacuous. Inspired by this observation, we propose the following condition that can incorporate this case of $\gamma = 1$, and will be shown to suffice to make expert distillation effective.

Definition 3.2 (Deterministic filter condition)**.** We say a POMDP $\mathcal{P}$ satisfies the *deterministic filter* condition if for each $h \geq 2$, the belief update operator under $\mathcal{P}$ satisfies that there exists a function $\psi_h : \mathcal{S} \times \mathcal{A} \times \mathcal{O} \to \mathcal{S}$ such that for any reachable $s_{h-1} \in \mathcal{S}$, $o_h \in \mathcal{O}$, $a_{h-1} \in \mathcal{A}$, $U_h(b^{s_{h-1}}; a_{h-1}, o_h) = b^{\psi_h(s_{h-1}, a_{h-1}, o_h)}$, where we define for any $s \in \mathcal{S}$, $b^s \in \Delta(\mathcal{S})$ and $b^s(s) = 1$ to be a one-hot vector. In addition, for $h = 1$, there exists a function $\psi_1 : \mathcal{O} \to \mathcal{S}$ such that for any reachable o_1, $B_1(\mu; o_1) = b^{\psi_1(o_1)}$, where $B_h(b; o_h) := \mathbb{P}^{\mathcal{P}}_{s_h \sim b}(s_h \,|\, o_h)$, $U_h(b; a_{h-1}, o_h) := \mathbb{P}^{\mathcal{P}}_{s_{h-1} \sim b}(s_h \,|\, a_{h-1}, o_{h-1})$ are the belief update operators, for which we defer the formal introduction to Appendix C.1.

Notably, this condition is weaker than and thus covers several known tractable classes of POMDPs with sample and computation efficiency guarantees [30, 14, 31, 72], for which we refer to our Figure 1 for an illustration and defer the detailed examples to Appendix E.

In light of the pitfall in Proposition 3.1, we will analyze the efficiency, *both* computationally and statistically, of expert distillation in Section 4, under Definition 3.2.

3.2 Privileged Value Learning: Asymmetric Actor-Critic

Asymmetric actor-critic [59] iterates between two main procedures as in standard actor-critic algorithms [34], policy *improvement* and policy *evaluation*. As its name suggests, its key difference from the standard actor-critic algorithm is that the algorithm maintains Q-value functions based on the state, while the policy receives only the history as input.

Policy evaluation. At iteration $t-1$, given the policy π^{t-1}, the algorithm estimates Q values in the form of $\{Q_h^{t-1}(\tau_h, s_h, a_h)\}_{h\in[H]}$, where we adopt the unbiased version such that Q-functions are conditioned on *both history and states* [4][2]. One key to achieving sample efficiency is adding some bonus terms in policy evaluation to encourage exploration, i.e., obtaining some optimistic Q-function estimates, for which we defer the detailed introduction to Section 4.

Policy improvement. At each iteration t, given the policy evaluation $\{Q_h^{t-1}(\tau_h, s_h, a_h)\}_{h\in[H]}$ for π^{t-1}, the vanilla asymmetric actor critic algorithm updates the policy according to the *sample-based* gradient estimation via some trajectories $\{o^k_{1:H}, s^k_{1:H}, a^k_{1:H}\}_{k\in[K]}$ sampled from π^{t-1}

$$\pi^t \leftarrow \textsc{Proj}_\Pi \left(\pi^{t-1} + \frac{\lambda_t}{K} \sum_{k\in[K]} \sum_{h\in[H]} \nabla_\pi \log \pi_h^{t-1}(a_h^k \,|\, \tau_h^k) Q_h^{t-1}(\tau_h^k, s_h^k, a_h^k) \right), \tag{3.2}$$

where λ_t is the step-size and $\textsc{Proj}_\Pi$ is the projection operator onto the space of Π. Here we point out the potential drawback of the vanilla algorithm as in [59, 4], where the key insight is that for each policy evaluation and update step, one roughly *only* performs the computation of order $\mathcal{O}(KH)$, while needing to collect K new episodes. Thus, the *sample complexity* will scale in the same order as the *computational complexity*, which is super-polynomial even for γ-observable POMDPs [23].

Proposition 3.3 (Inefficiency of vanilla asymmetric actor-critic)**.** Under tabular parameterization for the policy and value, the vanilla asymmetric actor-critic algorithm (Equation (3.2)) suffers from super-polynomial sample complexity for γ-observable POMDPs under standard hardness assumptions.

To address such issues, we first point out that a proximal policy optimization [63] type policy improvement yields a nice closed-form of

$$\pi_h^t(\cdot \,|\, \tau_h) \propto \pi_h^{t-1}(\cdot \,|\, \tau_h) \exp\left(\eta \mathbb{E}_{s_h \sim \boldsymbol{b}_h(\tau_h)} \left[Q_h^{t-1}(\tau_h, s_h, \cdot)\right]\right), \forall h \in [H], \tau_h \in \mathcal{T}_h, \tag{3.3}$$

where we recall $\boldsymbol{b}_h(\tau_h) \in \Delta(\mathcal{S})$ denotes the belief state. We defer the detailed derivation to Appendix E. Here we can see that to perform this closed-form update, one needs not only the critic Q^{t-1}, but also the explicit belief function $\boldsymbol{b}_h(\tau_h) \in \Delta(\mathcal{S})$ and $\eta > 0$ is the learning rate.

[1] Note that γ cannot be larger than 1 since according to Assumption C.8, $\gamma \leq \|\mathbb{O}_h\|_\infty \leq 1$.

[2] Note that as pointed out in [4], the original asymmetric actor-critic [59], which conditioned the Q-function only on states, yields a *biased* estimate of the policy gradient.

However, such an update presents two challenges: (1) It requires enumerating all possible τ_h, whose number scales exponentially with the horizon, making it computationally intractable; (2) An explicit belief function $\boldsymbol{b}_h$ is needed. Motivated by these two caveats, we propose to consider *finite-memory*-based policy and assume access to an approximate belief function $\{\boldsymbol{b}_h^{\text{apx}} : \mathcal{Z}_h \to \Delta(\mathcal{S})\}_{h\in[H]}$ (the learning for which will be made clear later). Correspondingly, the policy update is modified as:

$$\pi_h^t(\cdot \mid z_h) \propto \pi_h^{t-1}(\cdot \mid z_h) \exp\left(\eta \mathbb{E}_{s_h \sim \boldsymbol{b}_h^{\text{apx}}(z_h)}\left[Q_h^{t-1}(z_h, s_h, \cdot)\right]\right), \forall h \in [H], z_h \in \mathcal{Z}_h.$$

Then we develop and analyze one possible approach to learning such an approximate belief (c.f. Section 5). It is worth noting that the policy optimization algorithm we aim to develop and analyze does not depend on the specific algorithm of learning the approximate belief. Such a decoupled framework shall facilitate and enable more flexible algorithm designs, and can potentially incorporate the rich literature for learning approximate beliefs.

Remark 3.4 (Importance of belief learning for better sample complexity)**.** We remark that learning/accessing an approximate belief function is the key to updating the policy *for all* $z_h \in \mathcal{Z}_h$ instead of only those sampled ones. This allows our approach to perform *more computation* at each iteration, thus potentially making a more effective use of samples and finally achieving polynomial sample complexity, standing in sharp contrast to the vanilla asymmetric actor-critic algorithms.

4 Provably Efficient Expert Policy Distillation

We now focus on the provable correctness and efficiency of expert policy distillation, under the deterministic filter condition in Definition 3.2. We will defer all the proofs in this section to Appendix F. Definition 3.2 motivates us to consider only succinct policies that incorporate an auxiliary parameter representing the most recent state, as well as the most recent observations and actions. We consider policies that are the composition of two functions: at step h a function $g_h : \mathcal{S} \times \mathcal{A} \times \mathcal{O} \to \mathcal{S}$ that decodes the state based on the previous state, the most recent action, and the most recent observation, and a policy $\pi^E \in \Pi_{\mathcal{S}}$ that takes as input the current (decoded) latent state and outputs a distribution over actions.

Definition 4.1. We define a policy class Π^D as:

$$\Pi^D = \left\{\pi_h^E\left(g_h(s_{h-1}, a_{h-1}, o_h)\right) : g_h : \mathcal{S} \times \mathcal{A} \times \mathcal{O} \to \mathcal{S}, \pi_h^E : \mathcal{S} \to \Delta(\mathcal{A})\right\}_{h\in[H]},$$

where π^E stands for an arbitrary expert policy, and Π^D stands for the distilled policy class, and for $h = 1$, a_0, s_0 are some fixed dummy action and state. Intuitively, the distilled policy $\pi \in \Pi^D$ executes as follows: it firstly decodes the underlying states by applying $\{g_h\}_{h\in[H]}$ *recursively* along the history, and then takes actions using π^E based on the decoded states.

Our goal is to learn the two functions independently, that is, we want to learn an approximately optimal policy $\pi^E \in \Pi_{\mathcal{S}}$ with respect to the MDP $\mathcal{M}$ associated with the POMDP $\mathcal{P}$ by omitting the observations and observing the latent state, and for each step $h \in [H]$, a decoding function $g_h(s_{h-1}, a_{h-1}, o_h)$ such that the probability that we incorrectly decode a state-action-observation triplet over the trajectories induced by policy π^E is low.

Definition 4.2. Consider a POMDP $\mathcal{P}$ that satisfies Definition 3.2, and let ψ_h be the promised function that always correctly decodes a state-action-observation triplet into a latent state in Definition 3.2. Consider policy $\widetilde{\pi^E} = \{\pi^E(\psi(\cdot)) : \mathcal{S} \times \mathcal{A} \times \mathcal{O} \to \mathcal{A}\} \in \Pi^D$. We abuse notation and simply denote by $v^{\mathcal{P}}(\pi^E) = v^{\mathcal{P}}(\widetilde{\pi^E})$.

Lemma 4.3. Let $\mathcal{P} = (H, \mathcal{S}, \mathcal{A}, \mathcal{O}, \mathbb{T}, \mathbb{O}, \mu_1, r)$ be a POMDP that satisfies Definition 3.2, and consider a policy $\pi^E \in \Pi_{\mathcal{S}}$. Consider a set of decoding functions $\{g_h\}_{h\in[H]}$ such that, $\mathbb{P}^{\pi^E,\mathcal{P}}\left[\exists h \in [H] : g_h\left(s_{h-1}, a_{h-1}, o_h\right) \neq s_h\right] \leq \epsilon$. Consider the policy $\pi = \left\{\pi_h^E\left(g_h(\cdot)\right) : \mathcal{S} \times \mathcal{A} \times \mathcal{O} \to \Delta(\mathcal{A})\right\}_{h\in[H]}$ on POMDP $\mathcal{P}$, then: $v^{\mathcal{P}}(\pi) \geq v^{\mathcal{P}}(\widetilde{\pi^E}) - H\epsilon$.

We can use any off-the-shelf algorithm to learn the approximately optimal policy π^E for the associated MDP $\mathcal{M}$. Thus in the rest of the section, we focus on learning the decoding function $\{g_h\}_{h\in[H]}$. To be able to efficiently learn the decoding function, we model the access to the underlying state with an MDP that keeps track of the most recent pair of the action taken, and the observation received, as well as the two most recent states.

Theorem 4.4. Consider a POMDP $\mathcal{P}$ that satisfies Definition 3.2, a policy $\pi^E \in \Pi_{\mathcal{S}}$, and let $\{g_h\}_{h\in[H]}$ be the outcome of Algorithm 1 with $M = \frac{AOS+\log(H/\delta)}{\epsilon^2}$. Then with probability at least $1-\delta$, for each level $h \in [H]$:

$$\mathbb{P}^{\pi^E,\mathcal{P}}\ [\exists h \in [H] : g_h\left(s_{h-1}, a_{h-1}, o_h\right) \neq s_h] \leq \epsilon,$$

using $\text{POLY}(H, A, O, S, \frac{1}{\epsilon}, \log\left(\frac{1}{\delta}\right))$ episodes in time $\text{POLY}(H, A, O, S, \frac{1}{\epsilon}, \log\left(\frac{1}{\delta}\right))$.

The following is an immediate consequence of Lemma 4.3 and Theorem 4.4. Note that both the sample and computation complexities are *polynomial*, which is in stark contrast to the k-decodable POMDP case [14] (a special one covered by our Definition 3.2), for which the sample complexity is necessarily *exponential in* k when there is no privileged information [14]. In fact, thanks to privileged information, the complexities are only polynomial in horizon H even when the decodable length is unknown. For the benefits of privileged information in several other subclasses of problems, we refer to Table 1 for more details.

Theorem 4.5. Let $\mathcal{P}$ satisfy Definition 3.2 and consider any policy $\pi^E \in \Pi_{\mathcal{S}}$. Using both $\text{POLY}(H, A, O, S, \frac{1}{\epsilon}, \log\left(\frac{1}{\delta}\right))$ episodes and time $\text{POLY}(H, A, O, S, \frac{1}{\epsilon}, \log\left(\frac{1}{\delta}\right))$, we can learn a policy $\pi \in \Pi^D$ (see Definition 4.1) such that with probability at least $1-\delta$, $v^{\mathcal{P}}(\pi) \geq v^{\mathcal{P}}(\pi^E) - \epsilon$.

Extensions to the case with general function approximation. Due to the compatibility of our algorithm with supervised learning oracles, it can be readily generalized to the function approximation setting to handle large observation spaces. We defer the corresponding results to Appendix G.

5 Provable Asymmetric Actor-Critic with Approximate Belief Learning

Unlike most existing theoretical studies on provably sample-efficient partially observable RL [31, 21, 39], which directly learn an approximate *POMDP model* for planning near-optimal policies, we consider a general framework with two steps: firstly learning an approximate *belief function*, followed by adopting an *fully observable RL* subroutine on the belief state space.

5.1 Belief-Weighted Optimistic Asymmetric Actor-Critic

We now introduce our main algorithmic contribution to the privileged policy learning setting. Our algorithm is conceptually similar to the natural policy gradient (NPG) methods [64] in the fully observable setting, and is presented in Algorithm 2. To adapt NPG to the partially observable setting, we need a subroutine that takes the stored memory as input and outputs a belief about the latent state (c.f. $\{\boldsymbol{b}_h^{\text{apx}}\}_{h\in[H]}$). Additionally, similar to the fully observable setting, we include a subroutine to estimate the Q-function, which introduces additional challenges due to partial observability (see Appendix H). We summarize the performance of Algorithm 2 in the following theorem. Our algorithm decouples the planning from the belief learning and estimation of the Q-function.

Theorem 5.1 (Near-optimal policy)**.** Fix $\epsilon, \delta \in (0,1)$ and Π^L. Given a POMDP $\mathcal{P}$ and an approximate belief $\{\boldsymbol{b}_h^{\text{apx}} : \mathcal{Z}_h \to \Delta(\mathcal{S})\}_{h\in[H]}$, with probability $1-\delta$, Algorithm 2 can learn an approximately optimal policy $\pi^\star$ of $\mathcal{P}$ in the space of Π^L such that

$$v^{\mathcal{P}}(\pi^\star) \geq \max_{\pi\in\Pi^L} v^{\mathcal{P}}(\pi) - \mathcal{O}(\epsilon + H^2\epsilon_{\text{belief}}),$$

with sample complexity $\text{POLY}(S, A, O, H, \frac{1}{\epsilon}, \log\frac{1}{\delta})$ and time complexity $\text{POLY}(S, A, O, H, Z, \frac{1}{\epsilon}, \log\frac{1}{\delta})$, with ϵ_{belief} defined as the total variation distance $\epsilon_{\text{belief}} := \max_{h\in[H]} \max_{\pi\in\Pi^L} \mathbb{E}^{\mathcal{P}}_{\pi\in\Pi^L} \|\boldsymbol{b}_h(\tau_h) - \boldsymbol{b}_h^{\text{apx}}(z_h)\|_1$ and $Z := \max_h |\mathcal{Z}_h|$. Furthermore, if $\mathcal{P}$ is additionally γ-observable (Assumption C.8), then $\pi^\star$ is also a global optimal policy in the space of Π such that $v^{\mathcal{P}}(\pi^\star) \geq \max_{\pi\in\Pi} v^{\mathcal{P}}(\pi) - \mathcal{O}(\epsilon + H^2\epsilon_{\text{belief}})$, as long as $L \geq \widetilde{\Omega}(\gamma^{-4}\log(SH/\epsilon))$.

5.2 Learning Approximate Belief

On a high level, our belief-learning algorithm first learns an approximate POMDP model $\widehat{\mathcal{P}}$ by explicitly exploring the state space, and then performs a truncation on the learned transition and emission to ensure the *filter stability* under the *misspecified model*, followed by outputting its approximate belief function. Note that the key to achieving belief learning with *both* polynomial sample and time complexity is our explicit exploration on the state space, which relies on executing fully observable policies from an *MDP learning* subroutine. We remark that the belief function may also be learned even if the state space is only explored by partially observable policies, thus utilizing

only hindsight observability may be sufficient for this purpose [36]. However, for such exploration to be computationally tractable, one requires a computationally tractable *POMDP learning* subroutine, which is in fact our final goal. We refer to Algorithm 4 for a detailed discussion on why privileged state information is necessary for computational tractability instead of only hindsight observability. We summarize the guarantees in the next theorem and postpone the proof to Appendix H.

Theorem 5.2. Consider a γ-observable POMDP $\mathcal{P}$ and assume that $L \geq \widetilde{\Omega}(\gamma^{-4}\log(S/\epsilon))$ for an $\epsilon > 0$. We can construct an approximate belief $\boldsymbol{b}_h^{\text{apx}}$ using $\widetilde{\mathcal{O}}(\frac{S^2AH^2O+S^3AH^2}{\epsilon^2} + \frac{S^4A^2H^6O}{\epsilon\gamma^2})$ episodes in time $\text{POLY}\left(S, H, A, O, \frac{1}{\gamma}, \frac{1}{\epsilon}, \log\left(\frac{1}{\delta}\right)\right)$ such that with probability $1-\delta$, for any $\pi \in \Pi^L$ and $h \in [H]$ $\mathbb{E}_\pi^{\mathcal{P}} \|\boldsymbol{b}_h(\tau_h) - \boldsymbol{b}_h^{\text{apx}}(z_h)\|_1 \leq \epsilon$.

Theorem 5.2 shows that an approximate belief can be learned with both polynomial samples and time, which, combined with Theorem 5.1, yields the final polynomial sample and quasi-polynomial time guarantees below. In contrast to the case without privileged information [21, 23], the sample complexity is reduced from quasi-polynomial to polynomial for γ-observable POMDPs. Note that the computation complexity remains quasi-polynomial, which is known to be unimprovable even for planning [23]. The key to such an improvement, as pointed out in Remark 3.4, is the more practical update rule of actor-critic (in conjunction with our belief-weighted idea), which allows *more computation* at each iteration (instead of only performing computation at the *sampled* trajectories). This allows the total computation to remain quasi-polynomial, while the overall sample complexity becomes polynomial. A detailed comparison can be found in Table 1.

Theorem 5.3. Let $\mathcal{P}$ be a γ-observable POMDP, and consider $L \geq \widetilde{\Omega}(\gamma^{-4}\log(SH/\epsilon))$ for an $\epsilon > 0$. With probability at least $1-\delta$, we can learn a policy $\pi \in \Pi^L$ such that $V_1^\pi(s_1) \geq \text{argmax}_{\pi\in\Pi} V_1^\pi(s_1) - \epsilon$, using $\text{POLY}(S, H, 1/\epsilon, 1/\gamma, \log(1/\delta), O, A)$ episodes and in time $\text{POLY}(S, H, 1/\epsilon, \log(1/\delta), O^L, A^L)$.

6 Numerical Validation

To corroborate the provable efficiency of our algorithm, we perform numerical validation for both our paradigms. Here we mainly compare with two baselines, the vanilla asymmetric actor-critic [59], and asymmetric Q-learning [5], on two settings, POMDP under the deterministic filter condition and general POMDPs. We report the results in Table 2 and Figure 2, where our algorithms achieve the best performance. We defer the implementation details and discussions to Appendix I.

7 Extensions to Partially Observable MARL with Privileged Information

7.1 Privileged Policy Learning: Equilibrium Distillation

To understand how the deterministic filter condition for POMDPs can be extended for POSGs, we first note the following equivalent characterization for Definition 3.2.

Proposition 7.1. Definition 3.2 is equivalent to the following: for each $h \in [H]$, there exists $\phi_h : \mathcal{T}_h \to \mathcal{S}$ such that $\mathbb{P}^{\mathcal{P}}(s_h = \phi_h(\tau_h) \,|\, \tau_h) = 1$ for any reachable $\tau_h \in \mathcal{T}_h$.

This implies that at each step h, given the *entire* history information, the agent can uniquely decode the current latent state s_h. Therefore, we generalize this condition to POSGs by requiring that each agent can uniquely decode the current latent state s_h given the information it collects so far.

Definition 7.2 (Deterministic filter condition for POSGs)**.** We say a POSG $\mathcal{G}$ satisfies the deterministic filter condition if for each $i \in [n]$, $h \in [H]$, there exists $\phi_{i,h} : \mathcal{C}_h \times \mathcal{P}_{i,h} \to \mathcal{S}$ such that $\mathbb{P}^{\mathcal{G}}(s_h = \phi_{i,h}(c_h, p_{i,h}) \,|\, c_h, p_{i,h}) = 1$ for any reachable $c_h, p_{i,h}$.

Here we have required that each agent can decode the latent state through their own information *individually*. Therefore, one may wonder whether one can relax it so that only the *joint* history information of all the agents can decode the latent state. However, we point out in the following that it does not circumvent the computational hardness of POSG. Note that the computational hardness result can not be mitigated even with privileged state information, as the hardness we state here holds even for the planning problem with model knowledge, where once one has the model knowledge, one can simulate the RL problem with privileged information.

Proposition 7.3. Computing CCE in POSGs satisfying that there exists $\phi_h : \mathcal{C}_h \times \mathcal{P}_h \to \mathcal{S}$ such that $\mathbb{P}^{\mathcal{G}}(s_h = \phi_h(c_h, p_h) \,|\, c_h, p_h) = 1$ for any reachable c_h, p_h is still PSPACE-hard.

Learning multi-agent individual decoding functions with unilateral exploration. Similar to our framework for POMDPs, the framework we develop for POSGs is also decoupled into two steps, learning an expert fully observable equilibrium policy, and learning the decoding function, where the first step can be instantiated by any provable off-the-shelf algorithm for learning Markov games. The major difference from the framework for POMDPs lies in how to learn the decoding function. In Theorem J.1, we prove that the difference of the NE/CE/CCE gap between the expert policy and the translated student policy is bounded by decoding errors under policies from the *unilateral deviation* of the expert policy. Hence, given the expert policy π, the key algorithmic principle is to perform *unilateral exploration* for agent i to make sure the decoding function is accurate under policies (π_i', π_{-i}) for any π_i', keeping π_{-i} fixed. We refer the detailed algorithm to Algorithm 5, and present the guarantees for learning the decoding functions and the corresponding distilled policy for learning NE/CCE, while we defer the CE version to Theorem J.6.

Theorem 7.4 (Equilibria learning; combining Theorem J.1 and Theorem J.4)**.** Under Assumption C.9 and conditions of Definition 7.2, given an $\frac{\epsilon}{2}$-NE/CCE π for the associated (fully-observable) Markov game of $\mathcal{G}$, Algorithm 5 can learning the decoding function $\{\widehat{g}_{i,h}\}_{i\in[n],h\in[H]}$ such that with probability $1-\delta$, it is guaranteed that $\max_{u_i\in\Pi_i,j\in[n]} \mathbb{P}^{u_i\times\pi_{-i},\mathcal{G}}(s_h \neq \widehat{g}_{j,h}(c_h,p_{j,h})) \leq \frac{\epsilon}{4nH^2}$, for any $i\in[n], h\in[H]$ with both sample and computational complexity $\mathrm{POLY}(S,A,H,O,\frac{1}{\epsilon},\log\frac{1}{\delta})$. Consequently, policy $\pi^{\widehat{g}}$ distilled from π is an ϵ-NE/CCE of $\mathcal{G}$.

7.2 Privileged Value Learning: Asymmetric MARL with Approximate Belief Learning

For POMDPs, we have used finite-memory policies for computation efficiency. We adopt such a generalization to POSGs by defining the *compression* of the common information.

Definition 7.5 (Compressed approximate common information [48, 70, 43])**.** For each $h\in[H]$, given a set $\widehat{\mathcal{C}}_h$, we say $\mathrm{Compress}_h$ to be a compression function if $\mathrm{Compress}_h \in \{f:\mathcal{C}_h\to\widehat{\mathcal{C}}_h\}$. For each $c_h\in\mathcal{C}_h$, we denote $\widehat{c}_h := \mathrm{Compress}_h(c_h)$. We also require the compression function to satisfy the regularity that for each $h\in[H]$, there exists a function $\widehat{\Lambda}_{h+1}$ such that $\widehat{c}_{h+1} = \widehat{\Lambda}_{h+1}(\widehat{c}_h, \varpi_{h+1})$, for any $c_h\in\mathcal{C}_h$, $\varpi_{h+1}\in\Upsilon_{h+1}$, where we recall $c_{h+1} := c_h\cup\varpi_{h+1}$ in Assumption C.7.

Similar to the framework we developed for POMDPs in Section 5, we firstly develop the multi-agent RL algorithm based on some approximate belief, and then instantiate it with one provable approach for learning such an approximate belief.

Optimistic value iteration of POSGs with approximate belief. For POMDPs, the sufficient statistics for optimal decision-making is the posterior distribution over the state given history. However, for POSGs with information-sharing, as shown in [54, 53, 43], the sufficient statistics become the posterior distribution over the state *and the private information* given the common information, instead of only the state. Therefore, we consider the approximate belief in the form of $\widehat{P}_h : \widehat{\mathcal{C}}_h \to \Delta(\mathcal{P}_h\times\mathcal{S})$ for each $h\in[H]$, where we define the error compared with the ground-truth belief to be $\epsilon_{\text{belief}} := \max_{h\in[H]}\max_{\pi\in\Pi}\mathbb{E}_\pi^{\mathcal{G}}\sum_{s_h,p_h}|\mathbb{P}^{\mathcal{G}}(s_h,p_h\,|\,c_h) - \widehat{P}_h(s_h,p_h\,|\,\widehat{c}_h)|$, i.e., the *expected* total variation distance from the true one. We refer our algorithm to Algorithm 7, which is conceptually similar to the algorithm for POMDP, maintaining the asymmetric value function, and performing the policy update using the *belief-weighted* value function.

Theorem 7.6 (Equilibria learning; combining Theorem J.15 and Theorem J.16)**.** Fix $\epsilon,\delta\in(0,1)$. Under Assumption C.9 with probability $1-\delta$, Algorithm 7 can learn an $(\epsilon+H^2\epsilon_{\text{belief}})$-NE if $\mathcal{G}$ is zero-sum and $(\epsilon+H^2\epsilon_{\text{belief}})$-CE/CCE if $\mathcal{G}$ is general-sum with sample complexity $\mathcal{O}(\frac{H^4SAO\log(SAHO/\delta)}{\epsilon^2})$ and computation complexity $\mathrm{POLY}(S,(AO)^{\mathcal{O}(\gamma^{-4}\log(SH/\epsilon))},H,\frac{1}{\epsilon},\log\frac{1}{\delta})$.

Learning approximate belief with model truncation. The belief learning algorithm we design for POSGs is conceptually similar to that we designed for POMDPs, where the key to achieving both polynomial sample and computation complexity is still to firstly learn approximate models, i.e., transitions and emissions, and then carefully *truncate* its transition and emission to build the approximate belief, where we defer the detailed algorithm to Algorithm 8. In the following, we provide provable guarantees of both polynomial computation and sample complexity, which, combined with Theorem 7.6, leads to a final polynomial-samples and quasi-polynomial-time complexity result.

Theorem 7.7. For any $\epsilon>0$, under Assumption C.8, it holds that one can learn the approximate belief $\{\widehat{P}_h:\widehat{\mathcal{C}}_h\to\Delta(\mathcal{S}\times\mathcal{P}_h)\}_{h\in[H]}$ such that $\epsilon_{\text{belief}}\leq\frac{\epsilon}{H^2}$ with both polynomial sample complexity and computation complexity $\mathrm{POLY}(S,A,O,H,\frac{1}{\gamma},\frac{1}{\epsilon},\log\frac{1}{\delta})$ for all examples in Appendix C.4.

References

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

[2] O. M. Andrychowicz, B. Baker, M. Chociej, R. Jozefowicz, B. McGrew, J. Pachocki, A. Petron, M. Plappert, G. Powell, A. Ray, et al. Learning dexterous in-hand manipulation. *The International Journal of Robotics Research*, 39(1):3–20, 2020.

[3] R. Avalos, F. Delgrange, A. Nowe, G. Perez, and D. M. Roijers. The wasserstein believer: Learning belief updates for partially observable environments through reliable latent space models. In *The Twelfth International Conference on Learning Representations*, 2023.

[4] A. Baisero and C. Amato. Unbiased asymmetric reinforcement learning under partial observability. In *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems*, 2022.

[5] A. Baisero, B. Daley, and C. Amato. Asymmetric DQN for partially observable reinforcement learning. In *Uncertainty in Artificial Intelligence*, pages 107–117. PMLR, 2022.

[6] N. Brukhim, D. Carmon, I. Dinur, S. Moran, and A. Yehudayoff. A characterization of multiclass learnability, 2022.

[7] N. Brukhim, D. Carmon, I. Dinur, S. Moran, and A. Yehudayoff. A characterization of multiclass learnability. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 943–955. IEEE, 2022.

[8] S. Bubeck. Convex optimization: Algorithms and complexity. *Found. Trends Mach. Learn.*, 8(3-4):231–357, 2015.

[9] Q. Cai, Z. Yang, and Z. Wang. Reinforcement learning from partial observation: Linear function approximation with provable sample efficiency. In *International Conference on Machine Learning*, pages 2485–2522. PMLR, 2022.

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

[11] D. Chen, B. Zhou, V. Koltun, and P. Krähenbühl. Learning by cheating. In *Conference on Robot Learning*, pages 66–75. PMLR, 2020.

[12] F. Chen, Y. Bai, and S. Mei. Partially observable RL with b-stability: Unified structural condition and sharp sample-efficient algorithms. In *The Eleventh International Conference on Learning Representations*, 2023.

[13] X. Chen, Y. M. Mu, P. Luo, S. Li, and J. Chen. Flow-based recurrent belief state learning for pomdps. In *International Conference on Machine Learning*, pages 3444–3468. PMLR, 2022.

[14] Y. Efroni, C. Jin, A. Krishnamurthy, and S. Miryoosefi. Provable reinforcement learning with a short-term memory. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, and S. Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 5832–5850. PMLR, 2022.

[15] E. Even-Dar, S. M. Kakade, and Y. Mansour. The value of observation for monitoring dynamic systems. In M. M. Veloso, editor, *IJCAI 2007, Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India, January 6-12, 2007*, pages 2474–2479, 2007.

[16] E. Even-Dar, S. M. Kakade, and Y. Mansour. Online Markov decision processes. *Mathematics of Operations Research*, 34(3):726–736, 2009.

[17] J. Foerster, I. A. Assael, N. De Freitas, and S. Whiteson. Learning to communicate with deep multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 29, 2016.

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

[19] K. Fujii. Bayes correlated equilibria and no-regret dynamics. *arXiv preprint arXiv:2304.05005*, 2023.

[20] T. Gangwani, J. Lehman, Q. Liu, and J. Peng. Learning belief representations for imitation learning in pomdps. In *uncertainty in artificial intelligence*, pages 1061–1071. PMLR, 2020.

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

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

[23] N. Golowich, A. Moitra, and D. Rohatgi. Planning and learning in partially observable systems via filter stability. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 349–362, 2023.

[24] N. Golowich, A. Moitra, and D. Rohatgi. Exploration is harder than prediction: Cryptographically separating reinforcement learning from supervised learning. *arXiv preprint arXiv:2404.03774*, 2024.

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

[26] J. Guo, M. Chen, H. Wang, C. Xiong, M. Wang, and Y. Bai. Sample-efficient learning of pomdps with multiple observations in hindsight. In *The Twelfth International Conference on Learning Representations*, 2023.

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

[28] J. Hartline, V. Syrgkanis, and E. Tardos. No-regret learning in bayesian games. *Advances in Neural Information Processing Systems*, 28, 2015.

[29] H. Hu, A. Lerer, N. Brown, and J. Foerster. Learned belief search: Efficiently improving policies in partially observable settings. *arXiv preprint arXiv:2106.09086*, 2021.

[30] N. Jiang, A. Krishnamurthy, A. Agarwal, J. Langford, and R. E. Schapire. Contextual decision processes with low bellman rank are pac-learnable. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1704–1713. PMLR, 2017.

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

[32] C. Jin, A. Krishnamurthy, M. Simchowitz, and T. Yu. Reward-free exploration for reinforcement learning. In *International Conference on Machine Learning*, pages 4870–4879. PMLR, 2020.

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

[34] V. R. Konda and J. N. Tsitsiklis. Actor-critic algorithms. In *Advances in Neural Information Processing Systems*, pages 1008–1014, 2000.

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

[36] J. Lee, A. Agarwal, C. Dann, and T. Zhang. Learning in pomdps is sample-efficient with hindsight observability. In *International Conference on Machine Learning*, pages 18733–18773. PMLR, 2023.

[37] J. Lee, J. Hwangbo, L. Wellhausen, V. Koltun, and M. Hutter. Learning quadrupedal locomotion over challenging terrain. *Science robotics*, 5(47):eabc5986, 2020.

[38] S. Levine, C. Finn, T. Darrell, and P. Abbeel. End-to-end training of deep visuomotor policies. *The Journal of Machine Learning Research*, 17(1):1334–1373, 2016.

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

[40] Q. Liu, P. Netrapalli, C. Szepesvari, and C. Jin. Optimistic MLE: A generic model-based algorithm for partially observable sequential decision making. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 363–376, 2023.

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

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

[43] X. Liu and K. Zhang. Partially observable multi-agent RL with (quasi-)efficiency: the blessing of information sharing. In *International Conference on Machine Learning*, pages 22370–22419. PMLR, 2023.

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

[45] M. Lu, Y. Min, Z. Wang, and Z. Yang. Pessimism in the face of confounders: Provably efficient offline reinforcement learning in partially observable markov decision processes. In *The Eleventh International Conference on Learning Representations*, 2023.

[46] X. Lyu, A. Baisero, Y. Xiao, and C. Amato. A deeper understanding of state-based critics in multi-agent reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 9396–9404, 2022.

[47] X. Lyu, A. Baisero, Y. Xiao, B. Daley, and C. Amato. On centralized critics in multi-agent reinforcement learning. *Journal of Artificial Intelligence Research*, 77:295–354, 2023.

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

[49] G. B. Margolis, T. Chen, K. Paigwar, X. Fu, D. Kim, S. bae Kim, and P. Agrawal. Learning to jump from pixels. In *5th Annual Conference on Robot Learning*, 2021.

[50] T. Miki, J. Lee, J. Hwangbo, L. Wellhausen, V. Koltun, and M. Hutter. Learning robust perceptive locomotion for quadrupedal robots in the wild. *Science Robotics*, 7(62):eabk2822, 2022.

[51] D. Misra, M. Henaff, A. Krishnamurthy, and J. Langford. Kinematic state abstraction and provably efficient rich-observation reinforcement learning. In *International conference on machine learning*, pages 6961–6971. PMLR, 2020.

[52] P. Moreno, J. Humplik, G. Papamakarios, B. A. Pires, L. Buesing, N. Heess, and T. Weber. Neural belief states for partially observed domains. In *NeurIPS 2018 Workshop on Reinforcement Learning under Partial Observability*, 2018.

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

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

[55] H. Nguyen, A. Baisero, D. Wang, C. Amato, and R. Platt. Leveraging fully observable policies for learning under partial observability. In *Conference on Robot Learning*, 2022.

[56] H. Nguyen, B. Daley, X. Song, C. Amato, and R. Platt. Belief-grounded networks for accelerated robot learning under partial observability. In *Conference on Robot Learning*, pages 1640–1653. PMLR, 2021.

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

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

[59] L. Pinto, M. Andrychowicz, P. Welinder, W. Zaremba, and P. Abbeel. Asymmetric actor critic for image-based robot learning. *arXiv preprint arXiv:1710.06542*, 2017.

[60] S. Qiu, Z. Dai, H. Zhong, Z. Wang, Z. Yang, and T. Zhang. Posterior sampling for competitive rl: Function approximation and partial observation. *Advances in Neural Information Processing Systems*, 36, 2024.

[61] T. Rashid, M. Samvelyan, C. S. De Witt, G. Farquhar, J. Foerster, and S. Whiteson. QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning. In *International Conference on Machine learning*, pages 681–689, 2018.

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

[63] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[64] L. Shani, Y. Efroni, A. Rosenberg, and S. Mannor. Optimistic policy optimization with bandit feedback. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 8604–8613. PMLR, 2020.

[65] L. Shani, Y. Efroni, A. Rosenberg, and S. Mannor. Optimistic policy optimization with bandit feedback. In *International Conference on Machine Learning*, pages 8604–8613. PMLR, 2020.

[66] I. Shenfeld, Z.-W. Hong, A. Tamar, and P. Agrawal. Tgrl: An algorithm for teacher guided reinforcement learning. In *International Conference on Machine Learning*, pages 31077–31093. PMLR, 2023.

[67] M. Shi, Y. Liang, and N. Shroff. Theoretical hardness and tractability of pomdps in rl with partial online state information, 2024.

[68] S. M. Shortreed, E. Laber, D. J. Lizotte, T. S. Stroup, J. Pineau, and S. A. Murphy. Informing sequential clinical decision-making through reinforcement learning: an empirical study. *Machine learning*, 84:109–136, 2011.

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

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

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

[72] M. Uehara, A. Sekhari, J. D. Lee, N. Kallus, and W. Sun. Computationally efficient pac rl in pomdps with latent determinism and conditional embeddings. In *International Conference on Machine Learning*, pages 34615–34641. PMLR, 2023.

[73] A. Wang, A. C. Li, T. Q. Klassen, R. T. Icarte, and S. A. McIlraith. Learning belief representations for partially observable deep rl. In *International Conference on Machine Learning*, pages 35970–35988. PMLR, 2023.

[74] L. Wang, Q. Cai, Z. Yang, and Z. Wang. Represent to control partially observed systems: Representation learning with provable sample efficiency. In *The Eleventh International Conference on Learning Representations*, 2022.

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

[76] G. Yang, M. Liu, W. Hong, W. Zhang, F. Fang, G. Zeng, and Y. Lin. Perfectdou: Dominating doudizhu with perfect information distillation. *Advances in Neural Information Processing Systems*, 35:34954–34965, 2022.

[77] Y. Yang, Y. Jiang, J. Chen, S. E. Li, Z. Gu, Y. Yin, Q. Zhang, and K. Yu. Belief state actor-critic algorithm from separation principle for POMDP. In *2023 American Control Conference (ACC)*, pages 2560–2567. IEEE, 2023.

[78] S. Young, M. Gašić, B. Thomson, and J. D. Williams. Pomdp-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179, 2013.

[79] W. Zhan, M. Uehara, W. Sun, and J. D. Lee. PAC reinforcement learning for predictive state representations. In *The Eleventh International Conference on Learning Representations*, 2023.

Supplementary Materials for "Provable Partially Observable Reinforcement Learning with Privileged Information"

Contents

A Societal Impact

Our work is theoretical, and aimed at better understanding reinforcement learning under partial observability with privileged information. As such, we do not anticipate any direct positive or negative societal impact from this research.

B Related Work

Provable partial observable RL. While POMDPs are generally known to be both statistically hard [35] and computationally intractable [57], a productive line of research has identified several interesting structured subclasses of POMDPs that can be efficiently solved. [35] introduced the class of POMDPs in the rich-observation setting, where the observation space can be large and fully reveal the latent state, where sample efficient RL becomes possible [30, 51]. [14] introduced k-step decodable POMDPs, where the last k observation-action pairs uniquely determine the state, proposing polynomial sample algorithms (assuming k is a small constant). Beyond settings where the latent state can be *exactly* recovered, [31, 39] proposed weakly revealing POMDPs, where the observations are assumed to be informative enough. Under the weakly revealing condition (and its variant), there has been a fast-growing line of recent works developing sample-efficient RL algorithms for various settings, see e.g., [74, 12, 9, 45, 40, 79]. Notably, these algorithms are typically computationally inefficient, requiring access to an optimistic planning oracle for POMDPs. On a promising note, [22] showed that in observable POMDPs (see Assumption C.8), one can achieve quasi-polynomial time for planning the near-optimal policy, which further leads to provable RL algorithms [21, 23] with *both quasi-polynomial* samples and computation complexities.

Comparison with RL under hindsight observability. The closest line of research to ours are the recent theoretical studies for Hindsight Observable Markov Decision Processes (HOMDPs) [36], where the latent state is revealed at the end of the episode; see also subsequent related works in [26, 67] with different observation feedback models. These works focused purely on *sample efficiency*, and showed that polynomial sample complexity can be achieved without (or by further relaxing) aforementioned structural assumptions of the model (e.g., observability or decodability), in both tabular and/or function approximation cases. However, the algorithms (also) require an oracle for planning or even optimistic planning in a learned approximate POMDP, which are not computationally tractable in general. Indeed, without any structural assumption, learning the optimal policy in HOMDPs is computationally no easier than the planning problem, which thus remains PSPACE-hard. Meanwhile, even under the additional assumption of observability, it is still not clear if these algorithms can avoid computationally intractable oracles, since the approximate POMDP that [36] needs to do planning on at every iteration during learning can be quite different from the ground-truth model. For example, at the beginning of exploration when no enough samples are collected, or when there exist certain states that remain less explored during the entire learning process, the potentially *misspecified emission (and transition)* may break the observability (or other structural) assumptions made for the *ground-truth* POMDP. This makes that single iteration computationally intractable. In contrast, our focus is on better understanding practically inspired algorithmic paradigms, without computationally intractable oracles, which in practice often do have privileged state information *during* each episode (instead of only at the end).

Most related empirical works. Privileged information has been widely used in empirical partially observable RL, with two main types of approaches based on privileged *policy* and privileged *value* learning, respectively. For the former, one prominent example is expert distillation [11, 55, 49], also known as *teacher-student* learning [37, 50, 66], as we analyze in Section 4. For the latter, asymmetric actor-critic [59] represents one of the well-known examples, with other studies in [5, 2]. Learning privileged value functions (to improve the policies) has also been widely used in multi-agent RL, featured in centralized-training-decentralized-execution, see e.g., [44, 18, 61, 76]. Intriguingly, it was shown that if the privileged value function depends *only on* the state, the associated actor will cause bias [4, 46, 47]. This has thus necessitated the use of history/belief in asymmetric actor-critic, as in our Section 5. Notably, the framework in [73] exactly matches ours, where they exploited the privileged state information in training for *belief learning*, followed by policy optimization over the learned belief states. Indeed, many empirical works explicitly separate the procedures of explicit belief-state learning and planning [20, 56, 29, 13, 77] as we study in Section 5, oftentimes with

privileged state information to supervise the belief learning procedure with better sample efficiency [52, 3].

C Additional Preliminaries

C.1 Additional Preliminaries on POMDPs

Belief and approximate belief. Although in a POMDP, the agent cannot see the underlying state directly, it can still form the *belief* over the latent state by the historical observations and action.

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

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

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

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

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

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

We also define the approximate belief update using the most recent L-step history. For $1 \le h \le H$, we follow the notation of [22] and define

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

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

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

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

C.2 Additional Preliminaries for POSGs

Model. We use a general framework of partially observable stochastic games (POSGs) as the model for partially observable MARL. Formally, we define a POSG with n agents by a tuple $\mathcal{G} = (H, \mathcal{S}, \{\mathcal{A}_i\}_{i=1}^n, \{\mathcal{O}_i\}_{i=1}^n, \mathbb{T}, \mathbb{O}, \mu_1, \{r_i\}_{i=1}^n)$, where H denotes the length of each episode, $\mathcal{S}$ is the state space with $|\mathcal{S}| = S$, $\mathcal{A}_i$ denotes the action space for the i^{th} agent with $|\mathcal{A}_i| = A_i$. We denote by $a := (a_1, \cdots, a_n)$ the joint action of all the n agents, and by $\mathcal{A} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_n$ the joint action space with $|\mathcal{A}| = A = \prod_{i=1}^n A_i$. We use $\mathbb{T} = \{\mathbb{T}_h\}_{h \in [H]}$ to denote the collection of transition matrices, so that $\mathbb{T}_h(\cdot \,|\, s, a) \in \Delta(\mathcal{S})$ gives the probability of the next state if joint action a is taken at state s and step h. In the following discussions, for any given a, we treat $\mathbb{T}_h(a) \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{S}|}$ as a matrix, where each row gives the probability for the next state. We use μ_1 to denote the distribution of the initial state s_1, and $\mathcal{O}_i$ to denote the observation space for the i^{th} agent with $|\mathcal{O}_i| = O_i$. We denote by $o := (o_1, \ldots, o_n)$ the joint observation of all n agents, and by $\mathcal{O} := \mathcal{O}_1 \times \cdots \times \mathcal{O}_n$ with $|\mathcal{O}| = O = \prod_{i=1}^n O_i$. We use $\mathbb{O} = \{\mathbb{O}_h\}_{h \in [H+1]}$ to denote the collection of the joint emission matrices, so that $\mathbb{O}_h(\cdot \,|\, s) \in \Delta(\mathcal{O})$ gives the emission distribution over the joint observation space $\mathcal{O}$ at state s and step h. For notational convenience, we will at times adopt the matrix convention, where $\mathbb{O}_h$ is a matrix with rows $\mathbb{O}_h(\cdot \,|\, s_h)$. We also denote $\mathbb{O}_{i,h}(\cdot \,|\, s) \in \Delta(\mathcal{O}_i)$ as the marginalized

emission for the i^{th} agent. Finally, $r_i = \{r_{i,h}\}_{h\in[H]}$ is a collection of reward functions, so that $r_{i,h}(s_h, a_h)$ is the reward of the i^{th} agent given the state s_h and joint action a_h at step h.

Similar to a POMDP, in a POSG, the states are not observable to the agents, and each agent can only access its own individual observations. The game proceeds as follows. At the beginning of each episode, the environment samples s_1 from μ_1. At each step h, each agent i observes its own observation $o_{i,h}$, where $o_h := (o_{1,h}, \ldots, o_{n,h})$ is sampled jointly from $\mathbb{O}_h(\cdot \mid s_h)$. Then each agent i takes the action $a_{i,h}$ and receives the reward $r_{i,h}(s_h, a_h)$. After that the environment transitions to the next state $s_{h+1} \sim \mathbb{T}_h(\cdot \mid s_h, a_h)$. The current episode terminates once s_{H+1} is reached.

Information sharing, common and private information. Each agent i in the POSG maintains its own information, $\tau_{i,h}$, a collection of historical observations and actions at step h, namely, $\tau_{i,h} \subseteq \{o_1, a_1, o_2, \cdots, a_{h-1}, o_h\}$, and the collection of the history at step h is given by $\mathcal{T}_{i,h}$.

In many practical examples, agents may share part of the history with each other, which may introduce more structure in the game that leads to both sample and computation efficiency. The information sharing splits the history into *common/shared* and *private* information for each agent. The *common information* at step h is a subset of the joint history τ_h: $c_h \subseteq \{o_1, a_1, o_2, \cdots, a_{h-1}, o_h\}$, which is available to *all the agents* in the system, and the collection of the common information is denoted as $\mathcal{C}_h$ and we define $C_h = |\mathcal{C}_h|$. Given the common information c_h, each agent also has the private information $p_{i,h} = \tau_{i,h} \setminus c_h$, where the collection of the private information for the i^{th} agent is denoted as $\mathcal{P}_{i,h}$ and its cardinality as $P_{i,h}$. The joint private information at step h is denoted as p_h, where the collection of the joint private history is given by $\mathcal{P}_h = \mathcal{P}_{1,h} \times \cdots \times \mathcal{P}_{n,h}$ and the corresponding cardinality is $P_h = \prod_{i=1}^n P_{i,h}$. We allow c_h or $p_{i,h}$ to take the special value $\emptyset$ when there is no common or private information. In particular, when $\mathcal{C}_h = \{\emptyset\}$, the problem reduces to the general POSG without any favorable information structure; when $\mathcal{P}_{i,h} = \{\emptyset\}$, every agent holds the same history, and it reduces to a POMDP when the agents share a common reward function and the goal is usually to find the team-optimal solution.

Policies and value functions. We define a stochastic policy for the i^{th} agent at step h as:

$$\pi_{i,h} : \Omega_h \times \mathcal{P}_{i,h} \times \mathcal{C}_h \to \Delta(\mathcal{A}_i). \tag{C.1}$$

The corresponding policy class is denoted as $\Pi_{i,h}$. Hereafter, unless otherwise noted, when referring to *policies*, we mean the policies given in the form of (C.1), which maps the available information of the i^{th} agent, i.e., the private information together with the common information, to the distribution over her actions. Here $\omega_{i,h} \in \Omega_h$ is the random seed, and Ω_h is the random seed space, which is shared among agents. We further denote $\Pi_i = \times_{h\in[H]}\Pi_{i,h}$ as the policy space for agent i and Π as the joint policy space. As a special case, we define the space of *deterministic* policy as $\widetilde{\Pi}_i$, where $\widetilde{\pi}_i \in \widetilde{\Pi}_i$ maps the private information and common information to a *deterministic* action for the i^{th} agent and the joint space as $\widetilde{\Pi}$.

We will define π_i as a sequence of policies for agent i at all steps $h \in [H]$, i.e., $\pi_i = (\pi_{i,1}, \cdots, \pi_{i,H})$. A (potentially correlated) joint policy is denoted as $\pi = \pi_1 \odot \pi_2 \cdots \odot \pi_n \in \Pi$. A *product* policy is denoted as $\pi = \pi_1 \times \pi_2 \cdots \times \pi_n \in \Pi$ if the distributions of drawing each seed $\omega_{i,h}$ for different agents are independent.

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

Definition C.2 (Value function)**.** For each agent $i \in [n]$ and step $h \in [H]$, given common information c_h and joint policy $\pi = \{\pi_i\}_{i=1}^n \in \Pi$, the *value function conditioned on the common information* of agent i is defined as: $V_{i,h}^{\pi,\mathcal{G}}(c_h) := \mathbb{E}_\pi^{\mathcal{G}}\left[\sum_{h'=h+1}^{H+1} r_{i,h'}(o_{h'}) \,\middle|\, c_h\right]$, where the expectation is taken over the randomness from the model $\mathcal{G}$, policy π, and the random seeds. For any $c_{H+1} \in \mathcal{C}_{H+1}$: $V_{i,H+1}^{\pi,\mathcal{G}}(c_{H+1}) := 0$. From now on, we will refer to it as *value function* for short.

Another key concept in our analysis is the belief about the state *and* the private information conditioned on the common information among agents. Formally, at step h, given policies from 1 to $h-1$, we consider the common-information-based conditional belief $\mathbb{P}_h^{\pi_{1:h-1},\mathcal{G}}(s_h, p_h \mid c_h)$. This belief not only infers the current underlying state s_h, but also each agent's private information p_h. With the common-information-based conditional belief, the value function given in Definition C.2 has the

following recursive structure:

$$V_{i,h}^{\pi,\mathcal{G}}(c_h) = \mathbb{E}_{\pi}^{\mathcal{G}}[r_{i,h+1}(o_{h+1}) + V_{i,h+1}^{\pi,\mathcal{G}}(c_{h+1}) \,|\, c_h], \tag{C.2}$$

where the expectation is taken over the randomness of (s_h, p_h, a_h, o_{h+1}) given π. With this relationship, we can define the *prescription-value* function correspondingly, a generalization of the *action-value* function in Markov games and MDPs, as follows.

Definition C.3 (Prescription-value function)**.** At step h, given the common information c_h, joint policies $\pi = \{\pi_i\}_{i=1}^n \in \Pi$, and prescriptions $\{\gamma_{i,h}\}_{i=1}^n \in \Gamma_h$, the *prescription-value function conditioned on the common information and joint prescription* of the i^{th} agent is defined as:

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

where prescription $\gamma_{i,h} \in \Delta(\mathcal{A}_i)^{P_{i,h}}$ replaces the partial function $\pi_{i,h}(\cdot \,|\, \omega_{i,h}, c_h, \cdot)$ in the value function. From now on, we will refer to it as *prescription-value function* for short. With such a prescription function, agents can take actions purely based on their local private information [54, 53].

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

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

Definition C.4 (ϵ-approximate Nash equilibrium with information sharing)**.** For any $\epsilon \geq 0$, a product policy $\pi^\star \in \Pi$ is an ϵ-approximate Nash equilibrium of the POSG $\mathcal{G}$ with information sharing if:

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

Definition C.5 (ϵ-approximate coarse correlated equilibrium with information sharing)**.** For any $\epsilon \geq 0$, a joint policy $\pi^\star \in \Pi$ is an ϵ-approximate coarse correlated equilibrium of the POSG $\mathcal{G}$ with information sharing if:

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

Definition C.6 (ϵ-approximate correlated equilibrium with information sharing)**.** For any $\epsilon \geq 0$, a joint policy $\pi^\star \in \Pi$ is an ϵ-approximate correlated equilibrium of the POSG $\mathcal{G}$ with information sharing if:

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

where m_i is called *strategy modification* and $m_i = \{m_{i,h,c_h,p_{i,h}}\}_{h,c_h,p_{i,h}}$, with each $m_{i,h,c_h,p_{i,h}} : \mathcal{A}_i \to \mathcal{A}_i$ being a mapping from the action set to itself. The space of m_i is denoted as $\mathcal{M}_i$. The composition $m_i \diamond \pi_i$ will work as follows: at the step h, when the agent i is given c_h and $p_{i,h}$, the action chosen to be $(a_{1,h}, \cdots, a_{i,h}, \cdots, a_{n,h})$ will be modified to $(a_{1,h}, \cdots, m_{i,h,c_h,p_{i,h}}(a_{i,h}), \cdots, a_{n,h})$. Note that this definition follows from that in [69, 42, 33, 43] when there exists common information, and is a natural generalization of the definition in the normal-form game case [62]. Meanwhile, we denote $\mathcal{M}_i^{\text{gen}}$ to be the space of all possible strategy modification m_i if it conditions on any history information instead of only $c_h, p_{i,h}$. Similarly, we denote $\mathcal{M}_{\mathcal{S},i}$ to be the space of all possible strategy modification m_i if it only conditions on the current state.

C.2.1 Evolution of the Common and Private Information

Assumption C.7 (Evolution of common and private information)**.** We assume that common information and private information evolve over time as follows:

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

  $$\varpi_{h+1} = \chi_{h+1}(p_h, a_h, o_{h+1}), \tag{C.3}$$

  where χ_{h+1} is a fixed transformation. We use Υ_{h+1} to denote the collection of ϖ_{h+1} at step h.

- Private information evolves according to:

$$p_{i,h+1} = \xi_{i,h+1}(p_{i,h}, a_{i,h}, o_{i,h+1}), \tag{C.4}$$

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

Equation (C.3) states that the increment in the common information depends on the "new" information (a_h, o_{h+1}) generated between steps h and $h+1$ and part of the old information p_h. The incremental common information can be implemented by certain sharing and communication protocols among the agents. Equation (C.4) implies that the evolution of private information only depends on the newly generated private information $a_{i,h}$ and $o_{i,h+1}$. These evolution rules are standard in the literature [53, 54], specifying the source of common information and private information. Based on such evolution rules, we define $\{f_h\}_{h\in[H+1]}$ and $\{g_h\}_{h\in[H+1]}$, where $f_h : \mathcal{A}^h \times \mathcal{O}^h \to \mathcal{C}_h$ and $g_h : \mathcal{A}^h \times \mathcal{O}^h \to \mathcal{P}_h$ for $h \in [H+1]$, as the mappings that map the joint history to common information and joint private information, respectively.

C.3 Technical Assumptions

We now lay out several technical assumptions that can circumvent the known computational hardness of POMDP/POSGs, which may be used later for different approaches.

A key technical assumption is that the POMDPs/POSGs we consider satisfy an observability assumption, outlined below. This observability assumption allows us to use short memory policies to approximate the optimal policy, and yields quasi-polynomial-time complexity for both planning and learning in POMDPs/POSGs [22, 21, 43].

Assumption C.8 (γ-observability [15, 22, 21])**.** Let $\gamma > 0$. For $h \in [H]$, we say that the matrix $\mathbb{O}_h$ satisfies the γ-observability assumption if for each $h \in [H]$, for any $b, b' \in \Delta(\mathcal{S})$,

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

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

Additionally, for a POSG to be computationally tractable, certain information-sharing is necessary [43]. We thus make the following assumption as in [43].

Assumption C.9 (Strategy independence of beliefs [53, 27, 43])**.** Consider any step $h \in [H]$, any policy $\pi \in \Pi$, and any realization of common information c_h that has a non-zero probability under the trajectories generated by $\pi_{1:h-1}$. Consider any other policies $\pi'_{1:h-1}$, which also give a non-zero probability to c_h. Then, we assume that: for any such $c_h \in \mathcal{C}_h$, and any $p_h \in \mathcal{P}_h, s_h \in \mathcal{S}$, $\mathbb{P}_h^{\pi_{1:h-1},\mathcal{G}}(s_h, p_h \mid c_h) = \mathbb{P}_h^{\pi'_{1:h-1},\mathcal{G}}(s_h, p_h \mid c_h)$.

We provide examples satisfying this assumption in Appendix C.2, which include the fully-sharing structure as in [23, 60] as a special case. Finally, we also assume that common information and private information evolve overtime properly in Assumption C.7, as standard in [53, 54, 43], which covers the models considered in [23, 60, 41].

C.4 Strategy Independence of Belief and Examples

Here we take the examples from [43] to illustrate the generality of the information-sharing framework.

Example C.10 (One-step delayed sharing)**.** At any step $h \in [H+1]$, the common and private information are given as $c_h = \{o_{2:h-1}, a_{1:h-1}\}$ and $p_{i,h} = \{o_{i,h}\}$, respectively. In other words, the players share all the action-observation history until the previous step $h-1$, with only the new observation being the private information. This model has been shown useful for power control [1].

Example C.11 (State controlled by one controller with asymmetric delay sharing)**.** We assume there are 2 players for convenience. It extends naturally to n-player settings. Consider the case where the state dynamics are controlled by player 1, i.e., $\mathbb{T}_h(\cdot \mid s_h, a_{1,h}, a_{2,h}) = \mathbb{T}_h(\cdot \mid s_h, a_{1,h}, a'_{2,h})$ for all $s_h, a_{1,h}, a_{2,h}, a'_{2,h}, h$. There are two kinds of delay-sharing structures we could consider: **Case A:** the information structure is given as $c_h = \{o_{1,2:h}, o_{2,2:h-d}, a_{1,1:h-1}\}$, $p_{1,h} = \emptyset$, $p_{2,h} = \{o_{2,h-d+1:h}\}$, i.e., player 1's observations are available to player 2 instantly, while player 2's observations are available to player 1 with a delay of $d \geq 1$ time steps. **Case B:** similar to **Case A** but player 1's

Algorithm 1 Learning Decoding Function with Privileged Information

Require:

- POMDP $\mathcal{P}$,
- Expert policy $\pi^E \in \Pi_{\mathcal{S}}$,
- Number of samples per step M.

for each step $h \in [H]$ **do**

Collect M episodes $\left\{\left(s^{(i)}_{1:H+1}, o^{(i)}_{1:H}, a^{(i)}_{1:H}\right)\right\}_{i\in[M]}$ on POMDP $\mathcal{P}$ using policy π^E and let:

$$\widehat{D}_h := \left\{\left(s^{(i)}_{h-1}, a^{(i)}_{h-1}, o^{(i)}_{h}, s^{(i)}_{h}\right)\right\}_{i\in[M]}.$$

Define the decoding function g_h for level h as:

$$g_h(s_{h-1}, a_{h-1}, o_h) = \left\{s_h : (s_{h-1}, a_{h-1}, o_h, s_h) \in \widehat{D}_h\right\}$$

end for

return $\{g_h : \mathcal{S} \times \mathcal{A} \times \mathcal{O} \to \mathcal{S}\}_{h\in[H]}$

observation is available to player 2 with a delay of 1 step. The information structure is given as $c_h = \{o_{1,2:h-1}, o_{2,2:h-d}, a_{1,1:h-1}\}$, $p_{1,h} = \{o_{1,h}\}$, $p_{2,h} = \{o_{2,h-d+1:h}\}$, where $d \geq 1$. This kind of asymmetric sharing is common in network routing [58], where packages arrive at different hosts with different delays, leading to asymmetric delay sharing among hosts.

Example C.12 (Symmetric information game)**.** Consider the case when all observations and actions are available for all the agents, and there is no private information. Essentially, we have $c_h = \{o_{2:h}, a_{1:h-1}\}$ and $p_{i,h} = \emptyset$. We will also denote this structure as *fully sharing* hereafter.

Example C.13 (Information sharing with one-directional-one-step delay)**.** Similar to the previous cases, we also assume there are 2 players for ease of exposition, and the case can be generalized to multi-player cases straightforwardly. Similar to the one-step delay case, we consider the situation where all observations of player 1 are available to player 2, while the observations of player 2 are available to player 1 with one-step delay. All past actions are available to both players. That is, in this case, $c_h = \{o_{1,2:h}, o_{2,2:h-1}, a_{1:h-1}\}$, and player 1 has no private information, i.e., $p_{1,h} = \emptyset$, and player 2 has private information $p_{2,h} = \{o_{2,h}\}$.

Example C.14 (Uncontrolled state process)**.** Consider the case where the state transition does not depend on the actions, that is, $\mathbb{T}_h(\cdot \mid s_h, a_h) = \mathbb{T}_h(\cdot \mid s_h, a'_h)$ for any s_h, a_h, a'_h, h. Note that the agents are still coupled through the joint reward. An example of this case is the information structure where controllers share their observations with a delay of $d \geq 1$ time steps. In this case, the common information is $c_h = \{o_{2:h-d}\}$ and the private information is $p_{i,h} = \{o_{i,h-d+1:h}\}$. Such information structures can be used to model repeated games with incomplete information.

D Collection of Algorithms

Algorithm 2 Belief-Weighted Optimistic Asymmetric Actor-Critic with Privileged Information

Require:

- Subroutine T_Q that given policy π, outputs $\{\widetilde{Q}_h^\pi\}_{h\in[H]}$ that approximates $\{Q_h^{\pi,\mathcal{P}}\}_{h\in[H]}$,
- Subroutine $T_{\boldsymbol{b}}$ that outputs $\{\boldsymbol{b}_h^{\text{apx}}\}_{h\in[H]}$ that approximate $\{\boldsymbol{b}_h\}_{h\in[H]}$,
- Initial finite-memory policy $\pi^0 = \{\pi_h^0\}_{h\in[H]} \in \Pi^L$ for the POMDP $\mathcal{P}$, step-size η, and number of iterations T.

Ensure: A near-optimal policy

$\{\boldsymbol{b}_h^{\text{apx}}\}_{h\in[H]} \leftarrow T_{\boldsymbol{b}}(\mathcal{P})$.

for Iterations $t = 1 \dots, T$ **do**

$\{\widetilde{Q}_h^{\pi^t}\}_{h\in[H]} \leftarrow T_Q(\mathcal{P}, \pi^{t-1})$

Update the policy for each $a_h \in \mathcal{A}$, $z_h \in \mathcal{Z}_h$ as

$$\pi_h^t(a_h \mid z_h) \propto \pi_h^{t-1}(a_h \mid z_h) \cdot \exp\left(\eta \mathbb{E}_{s_h \sim \boldsymbol{b}_h^{\text{apx}}(z_h)}\left[\widetilde{Q}_h^{\pi^{t-1}}((z_h, s_h), a_h)\right]\right).$$

Denote $\pi^t = \{\pi_h^t\}_{h\in[H]}$

end for

return A policy uniform at random from set $\{\pi^t\}_{t\in[T]}$

Algorithm 3 Optimistic Q-function Estimation with Privileged Information

Require:

POMDP $\mathcal{P}$, policy $\pi : \mathcal{S} \to \mathcal{A}$, number of episodes M per step.

Ensure: Approximate Q-functions $\{\widetilde{Q}_h^\pi\}_{h\in[H]}$ (see Lemma H.2)

Initialize:

$$\widetilde{Q}_{H+1}^\pi(z_{H+1}, s_{H+1}) \leftarrow 0, \qquad \forall z_{H+1} \in \mathcal{Z}_{H+1}, s_{H+1} \in \mathcal{S}, a_{H+1} \in \mathcal{A}.$$

for step $h = H, \dots, 1$ **do**

Collect M trajectories using policy π and let $D_h = \{\tau^{(i)}\}_{i\in[M]}$ be the collected trajectories.

Compute empirical counts and define empirical distributions:

$$\widehat{\mathbb{T}}_h(s_{h+1} \mid s_h, a_h) = \frac{|\{\overline{\tau} \in D_h : (s'_h, a'_h, s'_{h+1}) = (s_h, a_h, s_{h+1})\}|}{|\{\overline{\tau} \in D_h : (s'_h, a'_h) = (s_h, a_h)\}|}$$

$$\widehat{\mathbb{O}}_h(o_h \mid s_h) = \frac{|\{\overline{\tau} \in D_h : (s'_h, o'_h) = (s_h, o_h, s_{h+1})\}|}{|\{\overline{\tau} \in D_h : s'_h = s_h\}|}$$

for each memory-state pair $(z_h, s_h) \in \mathcal{Z}_h \times \mathcal{S}$ **do**

$$\begin{aligned}\widetilde{Q}_h^\pi((z_h, s_h), a_h) = \min\Bigg(&H - h + 1, \mathbb{E}_{\substack{s_{h+1}\sim\widehat{\mathbb{T}}_h(s_h,a_h),\\ o_{h+1}\sim\widehat{\mathbb{O}}_{h+1}(s_{h+1})}}[\widetilde{V}_{h+1}^\pi(z_{h+1}, s_{h+1})] \\ &+ r(s_h, a_h) + H \cdot \min\left(2, C \cdot \sqrt{\frac{|\mathcal{S}|\log(1/\delta_1)}{\max(N_h(s_h, a_h), 1)}}\right) \\ &+ \mathbb{E}_{s_{h+1}\sim\widehat{\mathbb{T}}_h(s_h,a_h)} H \cdot \min\left(2, C \cdot \sqrt{\frac{|\mathcal{O}|\log(1/\delta_1)}{\max(N_{h+1}(s_{h+1}), 1)}}\right)\Bigg),\end{aligned}$$

where $\widetilde{V}_h^\pi(z_h, s_h) = \mathbb{E}_{a_h\sim\pi(z_h)}[\widetilde{Q}_h^\pi(z_h, a_h)]$

end for

end for

return $\{\widetilde{Q}_h^\pi\}_{h\in[H]}$

Algorithm 4 Approximate Belief Learning via Model Truncation with Privileged Information

Input: $\mathcal{P} = (H, \mathcal{S}, \mathcal{A}, \mathcal{O}, \{\mathbb{T}_h\}_{h\in[H]}, \{\mathbb{O}_h\}_{h\in[H]}, \mu_1, \{r_h\}_{h\in[H]})$, an MDP learning oracle MDP_Learning that efficiently learns an approximate optimal policy of an MDP

for $h \in [H], s_h \in \mathcal{S}$ **do**

$\widehat{r}_{h'}(s'_h, a'_h) \leftarrow \mathbb{1}[h' = h, s'_h = s_h]$ for any $(h', s'_h, a'_h) \in [H] \times \mathcal{S} \times \mathcal{A}$.

$\mathcal{M} \leftarrow (H, \mathcal{S}, \mathcal{A}, \{\mathbb{T}_h\}_{h\in[H]}, \mu_1, \{\widehat{r}_h\}_{h\in[H]})$ to be the MDP associated with $\mathcal{P}$

$\Psi(h, s_h) \leftarrow$ MDP_Learning($\mathcal{M}$)

Collect N trajectories by executing policy $\Psi(h, s_h)$ for first $h-1$ steps then take action a_h for each $a_h \in \mathcal{A}$ deterministically and denote the dataset $\{(s^i_h, o^i_h, a^i_h, s^i_{h+1})\}_{i\in[NA]}$

for $o_h, a_h, s_{h+1}\mathcal{O} \times \mathcal{A} \times \mathcal{S}$ **do**

$N_h(s_h) \leftarrow \sum_{i\in[NA]} \mathbb{1}[s^i_h = s_h]$

$N_h(s_h, a_h) \leftarrow \sum_{i\in[NA]} \mathbb{1}[s^i_h = s_h, a^i_h = a_h]$

$N_h(s_h, a_h, s_{h+1}) \leftarrow \sum_{i\in[NA]} \mathbb{1}[s^i_h = s_h, a^i_h = a_h, s^i_{h+1} = s_{h+1}]$

$N_h(s_h, o_h) \leftarrow \sum_{i\in[NA]} \mathbb{1}[s^i_h = s_h, o^i_h = o_h]$

$\widehat{\mathbb{T}}_h(s_{h+1} \mid s_h, a_h) \leftarrow \frac{N_h(s_h, a_h, s_{h+1})}{N_h(s_h, a_h)}$

$\widehat{\mathbb{O}}_h(o_h \mid s_h) \leftarrow \frac{N_h(s_h, o_h)}{N_h(s_h)}$

end for

end for

for $h \in [H]$ **do**

$\mathcal{S}^{\text{high}}_h \leftarrow \left\{s_h \in \mathcal{S} \mid \frac{N_h(s_h)}{NA} \le \epsilon\right\}$

end for

for $h, s_h, o_h, a_h, s_{h+1} \in [H] \times \mathcal{S}^{\text{high}}_h \times \mathcal{O} \times \mathcal{A} \times \mathcal{S}^{\text{high}}_h$ **do**

$\widehat{\mathbb{T}}^{\text{trunc}}_h(s_{h+1} \mid s_h, a_h) \leftarrow \widehat{\mathbb{T}}_h(s_{h+1} \mid s_h, a_h)$

$\widehat{\mathbb{O}}^{\text{trunc}}_h(o_h \mid s_h) \leftarrow \widehat{\mathbb{O}}_h(o_h \mid s_h)$

$\widehat{\mathbb{T}}^{\text{trunc}}_h(s^{\text{exit}} \mid s_h, a_h) \leftarrow 1 - \sum_{s'_{h+1}\in\mathcal{S}^{\text{high}}_{h+1}} \widehat{\mathbb{T}}_h(s'_{h+1} \mid s_h, a_h)$

$\widehat{\mathbb{O}}^{\text{trunc}}_h(o^{\text{exit}} \mid s^{\text{exit}}) \leftarrow 1$

end for

Let $\widehat{\mathcal{P}}^{\text{trunc}} := (H, \{\mathcal{S}^{\text{high}}_h \cup \{s^{\text{exit}}\}\}_{h\in[H]}, \mathcal{A}, \mathcal{O} \cup \{o^{\text{exit}}\}, \{\widehat{\mathbb{T}}^{\text{trunc}}_h\}_{h\in[H]}, \{\widehat{\mathbb{O}}^{\text{trunc}}_h\}_{h\in[H]}, \mu_1, \{r_h\}_{h\in[H]})$

Define $\{\widehat{\boldsymbol{b}}'^{,\text{trunc}}_h : \mathcal{Z}_h \to \Delta(\mathcal{S}^{\text{high}}_h \cup \{s^{\text{exit}}\})\}_{h\in[H]}$ to be the approximate belief w.r.t. $\widehat{\mathcal{P}}^{\text{trunc}}$

Define $\{\boldsymbol{b}^{\text{apx}}_h : \mathcal{Z}_h \to \Delta(\mathcal{S})\}_{h\in[H]}$ such that $\boldsymbol{b}^{\text{apx}}_h(z_h)(s_h) = \widehat{\boldsymbol{b}}'^{,\text{trunc}}_h(z_h)(s_h) + \frac{\widehat{\boldsymbol{b}}'^{,\text{trunc}}_h(z_h)(s^{\text{exit}})}{|\mathcal{S}^{\text{high}}_h|}$ for $s_h \in \mathcal{S}^{\text{high}}_h$ and 0 otherwise.

return $\{\boldsymbol{b}^{\text{apx}}_h\}_{h\in[H]}$

E Missing Details in Section 3

Proof of Proposition 3.1: We recall that $\boldsymbol{b}_h(\cdot)$ is the belief of the agent about the latent state, see Appendix C for details. Note that Equation (3.1) can be written as

$$\arg\min_{\pi\in\Pi} \sum_{h=1}^{H} \mathbb{E}^{\mathcal{P}}_{\tau_h\sim\pi'} \mathbb{E}_{s_h\sim\boldsymbol{b}_h(\tau_h)} \left[D_f(\pi^\star_h(\cdot \mid s_h) \mid \pi_h(\cdot \mid \tau_h))\right].$$

Therefore, for any $h \in [H]$ and τ_h such that $\mathbb{P}^{\pi',\mathcal{P}}(\tau_h) > 0$, we can optimize π separately for each $h \in [H]$ and τ_h as:

$$\widehat{\pi}^\star_h(\cdot \mid \tau_h) \in \operatorname*{argmin}_{q\in\Delta(\mathcal{A})} \mathbb{E}_{s_h\sim\boldsymbol{b}_h(\tau_h)} \left[D_f(\pi^\star_h(\cdot \mid s_h) \mid q)\right].$$

Now we are ready to construct the counter-example γ-observable POMDP $\mathcal{P}^\epsilon$ with $H = 1$, $\mathcal{S} = \{s^1, s^2\}$, $\mathcal{A} = \{a^1, a^2\}$, and $\mathcal{O} = \{o^1, o^2\}$. We let $\mu_1 = (\frac{1}{2-\gamma}, \frac{1-\gamma}{2-\gamma})$, $\mathbb{O}_1(o^1 \mid s^1) = 1$, and $\mathbb{O}_1(o^1 \mid s^2) = 1-\gamma$, $\mathbb{O}_1(o^2 \mid s^2) = \gamma$. Therefore, it is direct to see that $\mathbb{O}_1$ is exactly γ-observable. Most importantly, we choose $r_1(s^1, a^1) = 1$, $r_1(s^1, a^2) = 0$, and $r_1(s^2, a^1) = 0$, $r_1(s^2, a^2) = \epsilon$.

Therefore, given such a reward function, the fully observable expert policy is given by $\pi^\star_1(a^1 \mid s^1) = 1$ and $\pi^\star_1(a^2 \mid s^2) = 1$, i.e., choosing a^1 at state s^1 and a^2 at state s^2 deterministically. Meanwhile,

Algorithm 5 Learning Multi-Agent (Individual) Decoding Functions with Privileged Information (NE/CCE Version)

Require: Input:

- $\mathcal{G} = (H, \mathcal{S}, \mathcal{A}, \mathcal{O}, \{\mathbb{T}_h\}_{h\in[H]}, \{\mathbb{O}_h\}_{h\in[H]}, \mu_1, \{r_i\}_{i\in[n]})$
- $\pi \in \Pi_{\mathcal{S}}$,
- controller set $\{\mathcal{I}_h \subseteq [n]\}_{h\in[H]}$

for $h \in [H]$, $s_h \in \mathcal{S}$ **do**
  for $i \in [n]$ **do**
    $\widehat{r}_{i,h'}(s'_h, a'_h) \leftarrow \mathbb{1}[h' = h, s'_h = s_h]$ for any $h', s'_h, a'_h \in [H] \times \mathcal{S} \times \mathcal{A}$.
    Define $\mathcal{M}$ to be the MDP associated with $\mathcal{G}$, $\mathcal{M}(\pi_{-i})$ to be the MDP marginalized by π_{-i}
    $\Psi_i(h, s_h) \leftarrow$ MDP_Learning$(\mathcal{M}(\pi_{-i}), \widehat{r}_i)$
  end for
  Collect N trajectories by executing policy $\Psi_i(h, s_h) \times \pi_{-i}$ for first $h-1$ steps then take action a_h deterministically for each $a_h \in \mathcal{A}$ and denote the dataset $\{(s_h^{k,i}, o_h^{k,i}, a_h^{k,i}, s_{h+1}^{k,i})\}_{k\in[NA]}$ for each $i \in [n]$
  for $o_h, a_h, s_{h+1} \in \mathcal{O} \times \mathcal{A} \times \mathcal{S}$ **do**
    $N_h(s_h) \leftarrow \sum_{k\in[N], i\in[n]} \mathbb{1}[s_h^{k,i} = s_h]$
    $N_h(s_h, a_{\mathcal{T}_h,h}, s_{h+1}) \leftarrow \sum_{k\in[N], i\in[n]} \mathbb{1}[s_h^{k,i} = s_h, a_{\mathcal{T}_h,h}^{k,i} = a_{\mathcal{T}_h,h}, s_{h+1}^{k,i} = s_{h+1}]$
    $N_h(s_h, o_h) \leftarrow \sum_{k\in[N], i\in[n]} \mathbb{1}[s_h^{k,i} = s_h, o_h^{k,i} = o_h]$
    $\widehat{\mathbb{T}}_h(s_{h+1} \mid s_h, a_{\mathcal{T}_h,h}) \leftarrow \frac{N_h(s_h, a_{\mathcal{T}_h,h}, s_{h+1})}{N_h(s_h, a_{\mathcal{T}_h,h})}$
    $\widehat{\mathbb{O}}_h(o_h \mid s_h) \leftarrow \frac{N_h(s_h, o_h)}{N_h(s_h)}$
  end for
end for
Define $\widehat{\mathcal{G}} := (H, \mathcal{S}, \mathcal{A}, \mathcal{O}, \{\widehat{\mathbb{T}}_h\}_{h\in[H]}, \{\widehat{\mathbb{O}}_h\}_{h\in[H]}, \mu_1, \{r_i\}_{i\in[n]})$
Define $\widehat{g}_{j,h}(s_h \mid c_h, p_{j,h}) := \mathbb{P}^{\widehat{\mathcal{G}}}(s_h \mid c_h, p_{j,h})$ for each $j \in [n]$, $h \in [H]$, $c_h \in \mathcal{C}_h$, $p_{j,h} \in \mathcal{P}_{j,h}$
return $\{\widehat{g}_{j,h}\}_{j\in[n],h\in[H]}$

by our delicate construction, one can compute that the belief given observation o^1 ensures $\boldsymbol{b}_1(o) = \mu_1 = \mathrm{Unif}(\mathcal{S})$. Hence, the corresponding "distilled" partially observable policy under observation o^1 is given by

$$\begin{aligned}
\widehat{\pi}_1^\star(\cdot \mid o^1) &= \operatorname*{argmin}_{q\in\Delta(\mathcal{A})} \mathbb{E}_{s_1\sim\boldsymbol{b}_1(o)} \left[D_f(\pi_1^\star(\cdot \mid s_1) \mid q)\right] \\
&= \operatorname*{argmin}_{q\in\Delta(\mathcal{A})} \frac{D_f(\pi_1^\star(\cdot \mid s^1) \mid q) + D_f(\pi_1^\star(\cdot \mid s^2) \mid q)}{2} \\
&= \operatorname*{argmin}_{q\in\Delta(\mathcal{A})} \frac{f(1/q(a^1))q(a^1) + f(0)q(a^2) + f(0)q(a^1) + f(1/q(a^2))q(a^2)}{2} \\
&= \operatorname*{argmin}_{q\in\Delta(\mathcal{A})} \frac{f(0) + f(1/q(a^1))q(a^1) + f(1/q(a^2))q(a^2)}{2},
\end{aligned}$$

where the last step is due to $q \in \Delta(\mathcal{A})$. Now consider the function $g(x) = xf(1/x)$ for $x > 0$. It is direct to compute that $g'(x) = f(1/x) - \frac{f'(1/x)}{x}$, and $g''(x) = \frac{f''(1/x)}{x^3} \geq 0$ due to the convexity of the function f. Thus, we conclude that g is also convex. By Jensen's inequality, we have

$$\frac{f(1/q(a^1))q(a^1) + f(1/q(a^2))q(a^2)}{2} \geq f(2/(q(a^1) + q(a^2)))(q(a^1) + q(a^2))/2 = f(2)/2,$$

where the equality holds when $q(a^1) = q(a^2) = \frac{1}{2}$. This indicates that $\widehat{\pi}_1^\star(\cdot \mid o) = \mathrm{Unif}(\mathcal{A})$. On the other hand, it is direct to see that the optimal partially observable policy $\widetilde{\pi} \in \arg\max_{\pi\in\Pi} v^{\mathcal{P}}(\pi)$ satisfies $\widetilde{\pi}_1(a^1 \mid o^1) = 1$. Now we are ready to evaluate the optimality gap between $\widetilde{\pi}$ and $\widehat{\pi}^\star$ as follows

$$\begin{aligned}
v^{\mathcal{P}^\epsilon}(\widetilde{\pi}) - v^{\mathcal{P}^\epsilon}(\widehat{\pi}^\star) &= \mathbb{P}^{\mathcal{P}^\epsilon}(o^1)(V_1^{\widetilde{\pi},\mathcal{P}^\epsilon}(o^1) - V_1^{\widehat{\pi}^\star,\mathcal{P}^\epsilon}(o^1)) + \mathbb{P}^{\mathcal{P}^\epsilon}(o^2)(V_1^{\widetilde{\pi},\mathcal{P}^\epsilon}(o^2) - V_1^{\widehat{\pi}^\star,\mathcal{P}^\epsilon}(o^2)) \\
&\geq \mathbb{P}^{\mathcal{P}^\epsilon}(o^1)(V_1^{\widetilde{\pi},\mathcal{P}^\epsilon}(o^1) - V_1^{\widehat{\pi}^\star,\mathcal{P}^\epsilon}(o^1)),
\end{aligned}$$

Algorithm 6 Learning Multi-Agent (Individual) Decoding Functions with Privileged Information (CE Version)

Input: $\mathcal{G} = (H, \mathcal{S}, \mathcal{A}, \mathcal{O}, \{\mathbb{T}_h\}_{h\in[H]}, \{\mathbb{O}_h\}_{h\in[H]}, \mu_1, \{r_i\}_{i\in[n]})$ $\pi \in \Pi_{\mathcal{S}}$, controller set $\{\mathcal{I}_h \subseteq [n]\}$, procedure MDP_Learning$(\cdot,\cdot)$ that takes as input an MDP and a reward function and returns an approximate optimal policy.
for $h \in [H]$, $s_h \in \mathcal{S}$ **do**
  for $i \in [n]$ **do**
    $\widehat{r}_{i,h'}(s'_h, a'_h) \leftarrow \mathbb{1}[h' = h, s'_h = s_h]$ for any $h', s'_h, a'_h \in [H] \times \mathcal{S} \times \mathcal{A}$.
    Define $\mathcal{M}^{\text{extended}}(\pi)$ to be the *extended* MDP, which is defined in Definition J.5.
    $\Psi_i(h, s_h) \leftarrow$ MDP_Learning$(\mathcal{M}^{\text{extended}}(\pi), \widehat{r}_i)$
  end for
  Collect N trajectories by executing policy $\Psi_i(h, s_h) \times \pi_{-i}$ for first $h-1$ steps then take action a_h deterministically for each $a_h \in \mathcal{A}$ and denote the dataset $\{(s_h^{k,i}, o_h^{k,i}, a_h^{k,i}, s_{h+1}^{k,i})\}_{k\in[NA]}$ for each $i \in [n]$
  for $o_h, a_h, s_{h+1} \in \mathcal{O} \times \mathcal{A} \times \mathcal{S}$ **do**
    $N_h(s_h) \leftarrow \sum_{k\in[N],i\in[n]} \mathbb{1}[s_h^{k,i} = s_h]$
    $N_h(s_h, a_{\mathcal{T}_h,h}, s_{h+1}) \leftarrow \sum_{k\in[N],i\in[n]} \mathbb{1}[s_h^{k,i} = s_h, a_{\mathcal{T}_h,h}^{k,i} = a_{\mathcal{T}_h,h}, s_{h+1}^{k,i} = s_{h+1}]$
    $N_h(s_h, o_h) \leftarrow \sum_{k\in[N],i\in[n]} \mathbb{1}[s_h^{k,i} = s_h, o_h^{k,i} = o_h]$
    $\widehat{\mathbb{T}}_h(s_{h+1} \mid s_h, a_{\mathcal{T}_h,h}) \leftarrow \frac{N_h(s_h, a_{\mathcal{T}_h,h}, s_{h+1})}{N_h(s_h, a_{\mathcal{T}_h,h})}$
    $\widehat{\mathbb{O}}_h(o_h \mid s_h) \leftarrow \frac{N_h(s_h, o_h)}{N_h(s_h)}$
  end for
end for
Define $\widehat{\mathcal{G}} := (H, \mathcal{S}, \mathcal{A}, \mathcal{O}, \{\widehat{\mathbb{T}}_h\}_{h\in[H]}, \{\widehat{\mathbb{O}}_h\}_{h\in[H]}, \mu_1, \{r_i\}_{i\in[n]})$
Define $\widehat{g}_{j,h}(s_h \mid c_h, p_{j,h}) := \mathbb{P}^{\widehat{\mathcal{G}}}(s_h \mid c_h, p_{j,h})$ for each $j \in [n]$, $h \in [H]$, $c_h \in \mathcal{C}_h$, $p_{j,h} \in \mathcal{P}_{j,h}$
return $\{\widehat{g}_{j,h}\}_{j\in[n],h\in[H]}$

where the last step is due to the fact that $\widetilde{\pi}$ is the optimal policy and it can at least mimick $\widehat{\pi}^\star$ at observation o^2, leading to the fact that $V_1^{\widetilde{\pi},\mathcal{P}^\epsilon}(o^1) - V_1^{\widehat{\pi}^\star,\mathcal{P}^\epsilon}(o^1) \geq 0$. Now it is not hard to compute that

$$\mathbb{P}^{\mathcal{P}^\epsilon}(o^1) \geq 1 - \gamma.$$

Meanwhile, we can evaluate that

$$V_1^{\widetilde{\pi},\mathcal{P}^\epsilon}(o^1) = \frac{1}{2}, \quad V_1^{\widehat{\pi}^\star,\mathcal{P}^\epsilon}(o^1) = \frac{1+\epsilon}{4}$$

and correspondingly $V_1^{\widetilde{\pi},\mathcal{P}^\epsilon}(o^1) - V_1^{\widehat{\pi}^\star,\mathcal{P}^\epsilon}(o^1) = \frac{1-\epsilon}{4}$, implying that $v^{\mathcal{P}^\epsilon}(\widetilde{\pi}) - v^{\mathcal{P}^\epsilon}(\widehat{\pi}^\star) \geq \frac{(1-\gamma)(1-\epsilon)}{4}$. This concludes our proof. ■

Here to show the generality of our condition in Definition 3.2, we introduce the following well-known examples in the literature that satisfies Definition 3.2.

Example E.1 (Deterministic POMDP [31, 72])**.** We say a POMDP $\mathcal{P}$ is of deterministic transition if entries of matrices $\{\mathbb{T}_h\}_{h\in[H]}$ and the vector μ_1 are either 0 or 1. Note that we do not make any assumptions on the emission matrices.

Example E.2 (Block MDP [30])**.** We say a POMDP $\mathcal{P}$ is a block MDP if for any $h \in [H]$, $s_h, s'_h \in \mathcal{S}$, it holds that $\text{supp}(\mathbb{O}_h(\cdot \mid s_h)) \cap \text{supp}(\mathbb{O}_h(\cdot \mid s'_h)) = \emptyset$ when $s_h \neq s'_h$.

Example E.3 (k-step decodable POMDP [14])**.** We say a POMDP $\mathcal{P}$ is an k-step decodable POMDP if there exists an unknown decoder $\phi^\star = \{\phi_h^\star : \mathcal{Z}_h \to \mathcal{S}\}_{h\in[H]}$ such that for any $h \in [H]$ and reachable trajectory τ_h, $\mathbb{P}^{\mathcal{P}}(s_h = \phi_h^\star(z_h) \mid \tau_h) = 1$, where $\mathcal{Z}_h = (\mathcal{O} \times \mathcal{A})^{\min\{h-1,k-1\}} \times \mathcal{O}$, $z_h = ((o,a)_{k(h):h-1}, o_h)$, and $k(h) = \min\{h-k+1, 1\}$.

Finally, to understand how our condition can extend beyond known examples in the literature, we show that one can indeed allow the decoding length of Example E.3 to be unknown and arbitrary (instead of being a small known constant as in [14] to get provably efficient algorithms).

Algorithm 7 Optimistic Common-Information-Based Value Iteration with Privileged Information

Input: $\mathcal{G}, \epsilon_e, \{\widehat{P}_h : \widehat{\mathcal{C}}_h \to \Delta(\mathcal{S} \times \mathcal{P}_h)\}_{h\in[H]}$
for $k = 1, 2, \cdots, K$ **do**
  for $h \leftarrow H, H-1, \cdots, 1$ **do**
    for $\widehat{c}_h \in \widehat{\mathcal{C}}_h$ **do**
      $Q^{\text{high},k}_{i,h}(\widehat{c}_h, p_h, s_h, a_h) \leftarrow \min\left\{r_{i,h}(s_h, a_h) + b_h^{k-1}(s_h, a_h) + \mathbb{E}_{o_{h+1}\sim\widehat{\mathbb{J}}_h^{k-1}(\cdot\,|\,s_h,a_h)}\left[V^{\text{high},k}_{i,h+1}(\widehat{c}_{h+1})\right], H-h+1\right\}$
for $i \in [n]$
      $Q^{\text{low},k}_{i,h}(\widehat{c}_h, p_h, s_h, a_h) \leftarrow \max\left\{r_{i,h}(s_h, a_h) - b_h^{k-1}(s_h, a_h) + \mathbb{E}_{o_{h+1}\sim\widehat{\mathbb{J}}_h^{k-1}(\cdot\,|\,s_h,a_h)}\left[V^{\text{low},k}_{i,h+1}(\widehat{c}_{h+1})\right], 0\right\}$
for $i \in [n]$
      Define $Q^{\text{high},k}_{i,h}(\widehat{c}_h, \gamma_h) := \mathbb{E}_{s_h,p_h\sim\widehat{P}_h(\cdot,\cdot\,|\,\widehat{c}_h)}\mathbb{E}_{\{a_{j,h}\sim\gamma_{j,h}(\cdot\,|\,p_{j,h})\}_{j\in[n]}}\left[Q^{\text{high},k}_{i,h}(\widehat{c}_h, p_h, s_h, a_h)\right]$
for $i \in [n]$
      Define $Q^{\text{low},k}_{i,h}(\widehat{c}_h, \gamma_h) := \mathbb{E}_{s_h,p_h\sim\widehat{P}_h(\cdot,\cdot\,|\,\widehat{c}_h)}\mathbb{E}_{\{a_{j,h}\sim\gamma_{j,h}(\cdot\,|\,p_{j,h})\}_{j\in[n]}}\left[Q^{\text{low},k}_{i,h}(\widehat{c}_h, p_h, s_h, a_h)\right]$
for $i \in [n]$
      $\{\pi^k_{j,h}(\cdot\,|\,\cdot, \widehat{c}_h, \cdot)\}_{j\in[n]} \leftarrow$ Bayesian-CE/CCE$(\{Q^{\text{high},k}_{j,h}(\widehat{c}_h, \cdot)\}_{j\in[n]})$ (c.f. Appendix J.1)
      $V^{\text{high},k}_{i,h}(\widehat{c}_h) \leftarrow \mathbb{E}_{\omega_h}\left[Q^{\text{high},k}_{i,h}(\widehat{c}_h, \{\pi^k_{j,h}(\cdot\,|\,\omega_{j,h}, \widehat{c}_h, \cdot)\}_{j\in[n]})\right]$ for $i \in [n]$
      $V^{\text{low},k}_{i,h}(\widehat{c}_h) \leftarrow \mathbb{E}_{\omega_h}\left[Q^{\text{low},k}_{i,h}(\widehat{c}_h, \{\pi^k_{j,h}(\cdot\,|\,\omega_{j,h}, \widehat{c}_h, \cdot)\}_{j\in[n]})\right]$ for $i \in [n]$
    end for
  end for
  Execute π^k and get trajectory $(s^k_{1:H}, a^k_{1:H}, o^k_{1:H+1})$
  for $h \in [H], s_h \in \mathcal{S}, a_h \in \mathcal{A}, o_{h+1} \in \mathcal{O}$ **do**
    $N^k_h(s_h, a_h) \leftarrow \sum_{l\in[k]} \mathbb{1}[s^l_h = s_h, a^l_h = a_h]$
    $N^k_h(s_h, a_h, o_{h+1}) \leftarrow \sum_{l\in[k]} \mathbb{1}[s^l_h = s_h, a^l_h = a_h, o^l_{h+1} = o_{h+1}]$
    $\widehat{\mathbb{J}}^k_h(o_{h+1}\,|\,s_h, a_h) \leftarrow \frac{N^k_h(s_h,a_h,o_{h+1})}{N^k_h(s_h,a_h)}$
  end for
end for

Example E.4 (POMDP with arbitrary, unknown decodable length)**.** This example is similar to Example E.3, but the decoding length m is unknown and not necessarily a small constant.

Proof of Example E.1 & Example E.2 & Example E.3 & Example E.4: To see why those examples follow our Definition 3.2, it is indeed an immediate result of Proposition 7.1. ■

Proof of Proposition 3.3: Here we evaluate the computation complexity and sample complexity of each iteration t as follows

Sample complexity: The algorithm executes the policy π^t and collect K episodes, denoted as $\{o^k_{1:H}, s^k_{1:H}, a^k_{1:H}\}_{k\in[K]}$ sampled from π^{t-1}. Thus, the sample complexity of each iteration is $\mathcal{O}(K)$.

Computation complexity for policy evaluation: The policy evaluation of vanilla asymmetric actor critic algorithm is done by minimizing the Bellman error. In the finite horizon with tabular parameterization, it is equivalent to performing the following update for each $h \in [H]$ in a backward way and $k \in [K]$.

$$Q^t_h(\tau^k_h, s^k_h, a^k_h) \leftarrow (1-\alpha)Q^{t-1}_h(\tau^k_h, s^k_h, a^k_h) + \alpha\left(r_h(s^k_h, a^k_h) + \frac{1}{|\mathcal{J}(\tau^k_h, s^k_h, a^k_h)|}\sum_{j\in\mathcal{J}(\tau^k_h, s^k_h, a^k_h)} Q^t_{h+1}(\tau^j_{h+1}, s^j_{h+1}, a^j_{h+1})\right),$$

for some $\alpha \in (0, 1)$, where $\mathcal{J}(\tau^k_h, s^k_h, a^k_h) := \{j \in [K]\,|\,(\tau^j_h, s^j_h, a^j_h) = (\tau^k_h, s^k_h, a^k_h)\}$. Therefore, the computation complexity for this procedure is of $\text{POLY}(H, K)$.

Algorithm 8 Approximate Belief Learning for MARL with Privileged Information

Input: $\mathcal{G} = (H, \mathcal{S}, \mathcal{A}, \mathcal{O}, \{\mathbb{T}_h\}_{h\in[H]}, \{\mathbb{O}_h\}_{h\in[H]}, \mu_1, \{r_{i,h}\}_{i\in[n],h\in[H]})$, controller set $\{\mathcal{I}_h \subseteq [n]\}_{h\in[H]}$

for $h \in [H], s_h \in \mathcal{S}$ **do**

$\widehat{r}_{h'}(s'_h, a'_h) \leftarrow \mathbb{1}[h' = h, s'_h = s_h]$ for any $(h', s'_h, a'_h) \in [H] \times \mathcal{S} \times \mathcal{A}$.

$\mathcal{M} \leftarrow (H, \mathcal{S}, \mathcal{A}, \{\mathbb{T}_h\}_{h\in[H]}, \mu_1, \{\widehat{r}_h\}_{h\in[H]})$ to be the MDP associated with $\mathcal{P}$

$\Psi(h, s_h) \leftarrow$ MDP_Learning$(\mathcal{M})$

Collect N trajectories by executing policy $\Psi(h, s_h)$ for first $h-1$ steps then take action a_h for each $a_h \in \mathcal{A}$ deterministically and denote the dataset $\{(s^i_h, o^i_h, a^i_h, s^i_{h+1})\}_{i\in[NA]}$

for $o_h, a_h, s_{h+1} \in \mathcal{O} \times \mathcal{A} \times \mathcal{S}$ **do**

$N_h(s_h) \leftarrow \sum_{i\in[NA]} \mathbb{1}[s^i_h = s_h]$

$N_h(s_h, a_{\mathcal{I}_h,h}) \leftarrow \sum_{i\in[NA]} \mathbb{1}[s^i_h = s_h, a^i_{\mathcal{I}_h,h} = a_{\mathcal{I}_h,h}]$

$N_h(s_h, a_{\mathcal{I}_h,h}, s_{h+1}) \leftarrow \sum_{i\in[NA]} \mathbb{1}[s^i_h = s_h, a^i_{\mathcal{I}_h,h} = a_{\mathcal{I}_h,h}, s^i_{h+1} = s_{h+1}]$

$N_h(s_h, o_h) \leftarrow \sum_{i\in[NA]} \mathbb{1}[s^i_h = s_h, o^i_h = o_h]$

$\widehat{\mathbb{T}}_h(s_{h+1} \mid s_h, a_{\mathcal{I}_h,h}) \leftarrow \frac{N_h(s_h,a_h,s_{h+1})}{N_h(s_h,a_{\mathcal{I}_h,h})}$

$\widehat{\mathbb{O}}_h(o_h \mid s_h) \leftarrow \frac{N_h(s_h,o_h)}{N_h(s_h)}$

end for

end for

for $h \in [H]$ **do**

$\mathcal{S}^{\text{high}}_h \leftarrow \left\{s_h \in \mathcal{S} \mid \frac{N_h(s_h)}{NA} \le \epsilon\right\}$

end for

for $h, s_h, o_h, a_h, s_{h+1} \in [H] \times \mathcal{S}^{\text{high}}_h \times \mathcal{O} \times \mathcal{A} \times \mathcal{S}^{\text{high}}_h$ **do**

$\widehat{\mathbb{T}}^{\text{trunc}}_h(s_{h+1} \mid s_h, a_{\mathcal{I}_h,h}) \leftarrow \widehat{\mathbb{T}}_h(s_{h+1} \mid s_h, a_{\mathcal{I}_h,h})$

$\widehat{\mathbb{O}}^{\text{trunc}}_h(o_h \mid s_h) \leftarrow \widehat{\mathbb{O}}_h(o_h \mid s_h)$

$\widehat{\mathbb{T}}^{\text{trunc}}_h(s^{\text{exit}} \mid s_h, a_{\mathcal{I}_h,h}) \leftarrow 1 - \sum_{s'_{h+1}\in\mathcal{S}^{\text{high}}_{h+1}} \widehat{\mathbb{T}}_h(s'_{h+1} \mid s_h, a_{\mathcal{I}_h,h})$

$\widehat{\mathbb{O}}^{\text{trunc}}_h(o^{\text{exit}} \mid s^{\text{exit}}) \leftarrow 1$

end for

Let $\widehat{\mathcal{G}}^{\text{trunc}} := (H, \{\mathcal{S}^{\text{high}}_h \cup \{s^{\text{exit}}\}\}_{h\in[H]}, \mathcal{A}, \mathcal{O} \cup \{o^{\text{exit}}\}, \{\widehat{\mathbb{T}}^{\text{trunc}}_h\}_{h\in[H]}, \{\widehat{\mathbb{O}}^{\text{trunc}}_h\}_{h\in[H]}, \mu_1, \{r_{i,h}\}_{i\in[n],h\in[H]})$

Define $\{\widetilde{P}_h : \widehat{\mathcal{C}}_h \to \Delta(\mathcal{P}_h \times (\mathcal{S}^{\text{high}}_h \cup \{s^{\text{exit}}\}))\}_{h\in[H]}$ to be the approximate belief w.r.t. $\widehat{\mathcal{G}}^{\text{trunc}}$

Define $\{\widehat{P}_h : \widehat{\mathcal{C}}_h \to \Delta(\mathcal{S} \times \mathcal{P}_h)\}_{h\in[H]}$ such that $\widehat{P}_h(s_h, p_h \mid \widehat{c}_h) = \widetilde{P}_h(s_h, p_h \mid s_h) + \frac{\widetilde{P}_h(s^{\text{exit}}_h, p_h \mid \widehat{c}_h)}{|\mathcal{S}^{\text{high}}_h|}$ for $s_h \in \mathcal{S}^{\text{high}}_h$ and 0 otherwise.

return $\{b^{\text{apx}}_h\}_{h\in[H]}$

Computation complexity for policy improvement: For tabular parameterization, computing $\nabla \log \pi^{t-1}_h(a^k_h \mid \tau^k_h)$ takes $\mathcal{O}(1)$ computation. Hence the policy update in Equation (3.2) performs $\text{POLY}(H, K)$ computation.

Meanwhile, under the exponential time hypothesis, there is no polynomial time algorithm for even planning an ϵ-approximate optimal policy in γ-observable POMDPs [22]. This implies that the vanilla asymmetric actor-critic algorithm needs to take super-polynomial iterations to find an approximately optimal policy. This implies the corresponding sample complexity has to be super-polynomial.

Finally, we remark that even if we let the policy and Q function not depend on the entire history τ_h but only the finite memory z_h, the proof still holds. ■

Derivation for the closed-form update Equation (3.3). Note that the proximal policy optimization [63] update has the policy improvement as follows

$$\pi^t \leftarrow \arg\max_{\pi} \left\{ L^{t-1}(\pi) - \eta^{-1} \mathbb{E}^{\mathcal{P}}_{\pi^{t-1}} \left[\sum_{h\in[H]} \text{KL}(\pi_h(\cdot \mid \tau_h) \mid \pi^{t-1}_h(\cdot \mid \tau_h)) \right] \right\}, \tag{E.1}$$

where η is some learning rate and $L^{t-1}(\pi)$ is a first-order approximation of the expected accumulated rewards at π^{t-1}:

$$L^{t-1}(\pi) := v^{\mathcal{P}}(\pi^{t-1}) + \mathbb{E}^{\mathcal{P}}_{\pi^{t-1}}\left[\sum_{h\in[H]} \left\langle Q_h^{t-1}(\tau_h, s_h, \cdot), \pi_h(\cdot \mid \tau_h) - \pi_h^{t-1}(\cdot \mid \tau_h)\right\rangle\right].$$

By plugging $L^{t-1}(\pi)$ into Equation (E.1), with simple algebric manipulations, we prove that:

$$\pi_h^t(\cdot \mid \tau_h) \propto \pi_h^{t-1}(\cdot \mid \tau_h) \exp\left(\eta \mathbb{E}_{s_h \sim \boldsymbol{b}_h(\tau_h)}\left[Q_h^{t-1}(\tau_h, s_h, \cdot)\right]\right).$$

F Missing Details in Section 4

Proof of Lemma 4.3: The proof follows by the assumption that the total cumulative reward is at most H,

$$\begin{aligned}
v^{\mathcal{P}}(\pi) \geq & \mathbb{E}^{\mathcal{P}}_{\pi}\left[\left(\sum_{h\in[H]} r_h\right) \mathbb{1}[\forall h :\in [H] : g_h\left(s_{h-1}, a_{h-1}, o_h\right) = s_h]\right] \\
= & \mathbb{E}^{\mathcal{P}}_{\pi^E}\left[\left(\sum_{h\in[H]} r_h\right) \mathbb{1}[\forall h :\in [H] : g_h\left(s_{h-1}, a_{h-1}, o_h\right) = s_h]\right] \\
& - \mathbb{E}^{\mathcal{P}}_{\pi^E}\left[\left(\sum_{h\in[H]} r_h\right) \mathbb{1}[\exists h :\in [H] : g_h\left(s_{h-1}, a_{h-1}, o_h\right) \neq s_h]\right] \\
\geq & v^{\mathcal{P}}(\pi^E) - H\mathbb{P}^{\pi^E,\mathcal{P}}\left[\exists h :\in [H] : g_h\left(s_{h-1}, a_{h-1}, o_h\right) \neq s_h\right] \\
\geq & v^{\mathcal{P}}(\pi^E) - H\epsilon.
\end{aligned}$$

■

Proof of Theorem 4.4: For each step $h \in [H]$, we define D_h to be the distribution over latent state s_{h-1} at level $h-1$, taken action $a_{h-1} \in \mathcal{A}$ from π^E, latent state transitioned to $s_h \in \mathcal{S}$, and observation $o_{h+1} \sim \mathbb{O}_h(s_h)$. Formally, the probability that the sequence $(s_{h-1}, a_{h-1}, o_h, s_h)$ is sampled from D_h equals to,

$$D_h := \mathbb{P}^{\pi^E,\mathcal{P}}\left[s'_{h-1} = s_{h-1}, a'_{h-1} = a_{h-1}, o'_h = o_h, s'_h = s_h\right].$$

We first use union bound to decompose the probability that we incorrectly decode,

$$\begin{aligned}
\mathbb{P}^{\pi^E,\mathcal{P}}\left[\exists h \in [H] : g_h\left(s_{h-1}, a_{h-1}, o_h\right) \neq s_h\right] &\leq \sum_{h\in[H]} \mathbb{P}^{\pi^E,\mathcal{P}}\left[g_h\left(s_{h-1}, a_{h-1}, o_h\right) \neq s_h\right] \\
&= \sum_{h\in[H]} \mathbb{P}_{(s_{h-1},a_{h-1},o_h,s_h)\sim D_h}\left[g_h\left(s_{h-1}, a_{h-1}, o_h\right) \neq s_h\right]. \qquad \text{(F.1)}
\end{aligned}$$

For each $h \in [H]$, we can use M episodes to collect M samples from distribution D_h. Denote the set of collected samples by $\widehat{D}_h^M$. We define the decoding g_h for level $h \in [H]$ as follows:

$$g_h(s_{h-1}, a_{h-1}, o_{h-1}) = \{s_h \mid (s_{h-1}, a_{h-1}, o_{h-1}, s_h) \in \widehat{D}_h^M\}.$$

Observe that by Definition 3.2, $\{s_{h+1} \mid (s_h, a_h, o_h, s_{h+1}) \in \widehat{D}_h^M\}$ is either the empty set or contains only a single elements, in which case, it is true that $g_h(s_h, a_h, o_h) = \psi_h(s_h, a_h, o_h)$ (ψ is the real decoding function, see Definition 3.2). Moreover we abuse notation and let $\widetilde{D}_h^M$ be the empirical distribution induced by samples in $\widehat{D}_h^M$. Thus with probability at least $1 - \frac{\delta}{H}$ and setting $M = \Theta\left(\frac{|\mathcal{A}|\cdot|\mathcal{O}|\cdot|\mathcal{S}|+\log(H/\delta)}{\epsilon^2}\right)$ for each level $h \in [H]$ using result by [10]:

$$\begin{aligned}
\mathbb{P}^{\pi^E,\mathcal{P}}\left[g_h\left(s_{h-1},a_{h-1},o_h\right)\neq s_h\right] =& \mathbb{P}^{\pi^E,\mathcal{P}}\left[g_h\left(s_{h-1},a_{h-1},o_h\right)=\emptyset\right] \\
=& \mathbb{P}_{(s_{h-1},a_{h-1},o_h,s_h)\sim D_h}\left[(s_{h-1},a_{h-1},o_h,s_h)\notin \widehat{D}_h\right] \\
=& \sum_{u\in\mathrm{supp}(D_h)} \Pr_{u'\sim D_h}[u=u']\mathbb{1}[u\notin \mathrm{supp}(\widetilde{D}_h^M)] \\
\leq& d_{TV}(D_h,\widetilde{D_h}^M) \\
\leq& \epsilon.
\end{aligned}$$

Thus, by union-bound, with probability at least $1-\delta$ we have that for each level $h\in[H]$,

$$\sum_{h\in[H]} \mathbb{P}_{(s_{h-1},a_{h-1},o_h,s_h)\sim D_h}\left[g_h\left(s_{h-1},a_{h-1},o_h\right)\neq s_h\right]\leq\epsilon,$$

which in combination with Equation (F.1) concludes the proof. Finally we note that we used a total of $\Theta\left(H\cdot\frac{|\mathcal{A}|\cdot|\mathcal{O}|\cdot|\mathcal{S}|+\log(H/\delta)}{\epsilon^2}\right)$ episodes from the POMDP, and the computational time was $\mathrm{POLY}(H,|\mathcal{A}|,|\mathcal{O}|,|\mathcal{S}|,\frac{1}{\epsilon},\log\left(\frac{1}{\delta}\right))$.. ■

G Provably Efficient Expert Policy Distillation with Function Approximation

We now turn our attention to the rich-observation setting under the deterministic filter condition. Definition 3.2 motivates us to consider only succinct policies that incorporate an auxiliary parameter representing the most recent state, as well as the most recent observations and actions. To handle the large observation space, we further assume that for each level $h\in[H]$, the agent selects a decoding function g_h from a family of multi-class classifiers $\mathcal{F}_h\subset\{\mathcal{S}\times\mathcal{A}\times\mathcal{O}\to\mathcal{S}\}$. For the function class $\mathcal{F}_h$ we make the standard realizability assumption. We formally summarize our assumptions in Assumption G.1

Assumption G.1. We consider a POMDP that satisfies Definition 3.2. In addition, to derive learning algorithms that do not dependent on $|\mathcal{O}|$, for each level $h\in[H]$ we assume that we have access to a class of functions $\mathcal{F}_h:\mathcal{S}\times\mathcal{A}\times\mathcal{O}\to\mathcal{S}$ such that the perfect decoding function $\psi_h\in\mathcal{F}_h$.

We aim for our final bounds to depend on a complexity measure of the function class $\mathcal{F}$ rather than the cardinality of the observation space $\mathbb{O}$. We utilize the Daniely and Shalev-Shwartz Dimension (DS Dimension) (Theorem G.2), which characterizes PAC learnability for multi-classification [6]. Defining the DS dimension is outside the scope of our paper; we direct interested readers to [6] for further details. For intuition, readers can think of the DS Dimension as a certificate of PAC learnability without loss of intuition.

Theorem G.2 (Theorem 1 in [6])**.** Consider a family of multi-class classifiers $\mathcal{F}$ that map features in space $x\in\mathcal{X}$ to labels in space $y\in\mathcal{Y}$. Moreover, assume there is a joint probability distribution D over features in $\mathcal{X}$ and labels in $\mathcal{Y}$, and that there exists $g^*\in\mathcal{F}$ such that for each $(x,y)\in\mathrm{supp}(D)$, $g^*(x)=y$. Given n samples from D, there exists an algorithm that with probability at least $1-\delta$ outputs $\widetilde{g}\in\mathcal{F}$ s.t.

$$\mathbb{P}_{(x,y)\sim D}[\widetilde{g}(x)\neq y]\leq\widetilde{\mathcal{O}}\left(\frac{d_{DS}(\mathcal{F})^{3/2}+\log\left(\frac{1}{\delta}\right)}{n}\right).$$

We are now ready to present the main theorem of this section.

Theorem G.3. Consider a POMDP $\mathcal{P}$ that satisfies Definition 3.2, a policy $\pi^E\in\Pi_{\mathcal{S}}$, and let $\{\mathcal{F}_h\subseteq\{\mathcal{S}\times\mathcal{A}\times\mathcal{O}\to\mathcal{S}\}\}_{h\in[H]}$ be the decoding function class, and $\psi_h\in\mathcal{F}_h$ for each $h\in[H]$, i.e., $\{\mathcal{F}_h\}_{h\in[H]}$ is realizable. Then given access to the classification oracle of [7], there exists an algorithm learning the decoding function $\{g_h\}_{h\in[H]}$ such that with probability at least $1-\delta$, for each level $h\in[H]$:

$$\mathbb{P}^{\pi^E,\mathcal{P}}\left[\exists h\in[H]:g_h\left(s_{h-1},a_{h-1},o_h\right)\neq s_h\right]\leq\epsilon,$$

using $\mathcal{O}\left(\frac{H^2\left(\max_{h\in[H]} d_{DS}^{3/2}(\mathcal{F}_h)+\log\left(\frac{1}{\delta}\right)\right)}{\epsilon}\right)$ episodes, where $d_{DS}(\mathcal{F}_h)$ is the the complexity measure of $\mathcal{F}_h$.

Proof of Theorem G.3:

For each level $h \in [2, H]$, we define D_h to be the distribution over latent state s_{h-1} at level $h-1$, taken action $a_{h-1} \in \mathcal{A}$ from π^E, latent state transitioned to $s_h \in \mathcal{S}$, and hallucinated observation $o_{h+1} \sim T_h(s_h)$. Formally, the probability that the sequence $(s_{h-1}, a_{h-1}, o_h, s_h)$ is sampled from D_h equals to,

$$D_h := \mathbb{P}^{\pi^E,\mathcal{P}}\left[s'_{h-1} = s_{h-1}, a'_{h-1} = a_{h-1}, o'_h = o_h, s'_h = s_h\right].$$

We first use union bound to decompose the misclassification error,

$$\begin{aligned}\mathbb{P}^{\pi^E,\mathcal{P}}\left[\exists h \in [H] : \widetilde{g}_h\left(s_{h-1}, a_{h-1}, o_h\right) \neq s_h\right] &\leq \sum_{h\in[H]} \mathbb{P}^{\pi^E,\mathcal{P}}\left[\widetilde{g}_h\left(s_{h-1}, a_{h-1}, o_h\right) \neq s_h\right] \\ &= \sum_{h\in[H]} \mathbb{P}_{(s_{h-1},a_{h-1},o_h,s_h)\sim D_h}\left[\widetilde{g}_h\left(s_{h-1}, a_{h-1}, o_h\right) \neq s_h\right]. \end{aligned} \tag{G.1}$$

For each $h \in [H]$, we can use $\widetilde{\mathcal{O}}\left(\frac{H}{\epsilon}\left(d_{DS}(\mathcal{F}_h)^{3/2} + \log\left(\frac{H}{\delta}\right)\right)\right) = \widetilde{\mathcal{O}}\left(\frac{H}{\epsilon}\cdot\left(\max_{h\in[H]} d_{DS}(\mathcal{F}_h)^{3/2} + \log\left(\frac{1}{\delta}\right)\right)\right)$ episodes to collect $\widetilde{\mathcal{O}}\left(\frac{H}{\epsilon}\cdot\left(\max_{h\in[H]} d_{DS}(\mathcal{F}_h)^{3/2} + \log\left(\frac{1}{\delta}\right)\right)\right)$ samples from distribution D_h. Hence by Theorem G.2, with probability at least $1 - \frac{\delta}{H}$ we have that

$$\mathbb{P}_{(s_{h-1},a_{h-1},o_h,s_h)\sim D_h}\left[\widetilde{g}_h\left(s_{h-1}, a_{h-1}, o_h\right) \neq s_h\right] \leq \frac{\epsilon}{H}.$$

Thus, by union bound, with probability at least $1-\delta$, using a total of $\widetilde{\mathcal{O}}\left(\frac{H^2}{\epsilon}\cdot\left(\max_{h\in[H]} d_{DS}(\mathcal{F}_h)^{3/2} + \log\left(\frac{1}{\delta}\right)\right)\right)$ episodes we have that,

$$\sum_{h\in[H]} \mathbb{P}_{(s_{h-1},a_{h-1},o_h,s_h)\sim D_h}\left[\widetilde{g}_h\left(s_{h-1}, a_{h-1}, o_h\right) \neq s_h\right] \leq \epsilon,$$

which in combination with Equation (G.1) concludes the proof. ■

H Missing Details in Section 5

Proof of Theorem 5.1:

Let $\pi^* \in \operatorname{argmax}_{\pi\in\Pi^L} V_1^{\pi}(s_1)$. We first note the following equation,

$$\frac{1}{T}\sum_{t\in[T]} V_1^{\pi^t}(s_1) = V_1^{\pi^*}(s_1) + \frac{1}{T}\sum_{t\in[T]}\left(\widetilde{V}_1^{\pi^t}(s_1) - V_1^{\pi^*}(s_1)\right) + \frac{1}{T}\sum_{t\in[T]}\left(V_1^{\pi^t}(s_1) - \widetilde{V}_1^{\pi^t}(s_1)\right). \tag{H.1}$$

We make use of the extended performance difference lemma:

Lemma H.1 (Lemma 1 in [64])**.** For any pair of policies $\pi = \{\pi_h\}_{h\in[H]}, \pi' = \{\pi'_h\}_{h\in[H]}$, and approximation of the Q-function of policy π, we have that:

$$\begin{aligned}&\widetilde{V}_1^{\pi}(s_1) - V_1^{\pi'}(s_1) \\ &= \sum_{h\in[H]} \mathbb{E}_{\overline{\tau}_h\sim\pi'}\left[\left\langle \widetilde{Q}_h^{\pi}((z_h, s_h), \cdot), \pi_h(\cdot \mid z_h, s_h) - \pi'_h(\cdot \mid z_h, s_h)\right\rangle\right] \\ &\quad + \sum_{h\in[H]} \mathbb{E}_{\overline{\tau}_h\sim\pi'}\left[\widetilde{Q}_h^{\pi}((z_h, s_h), a_h) - \mathbb{E}_{\substack{s_{h+1}\sim\mathbb{T}_h(\cdot\mid s_h,a_h),\\ o_{h+1}\sim\mathbb{O}_{h+1}(\cdot\mid s_{h+1})}}\left[r_h(s_h, a_h) + \widetilde{V}_{h+1}^{\pi}(z_{h+1}, s_{h+1})\right]\right],\end{aligned}$$

where $\widetilde{V}_h^{\pi}(z_h, s_h) = \mathbb{E}_{a_h\sim\pi(z_h)}[\widetilde{Q}_h^{\pi}(z_h, a_h)]$.

Setting $\pi = \pi^t$, and $\pi' = \pi^*$, the above formulation is simplified to[3],

$$
\begin{aligned}
&\widetilde{V}_1^{\pi^t}(s_1) - V_1^{\pi^*}(s_1) \\
&= \sum_{h\in[H]} \mathbb{E}_{\overline{\tau}_h\sim\pi^*}\left[\left\langle \widetilde{Q}_h^{\pi^t}((z_h,s_h),\cdot), \pi_h^t(\cdot \mid z_h,s_h) - \pi_h^*(\cdot \mid z_h,s_h)\right\rangle\right] \\
&\quad + \sum_{h\in[H]} \mathbb{E}_{\overline{\tau}_h\sim\pi^*}\left[\widetilde{Q}_h^{\pi}((z_h,s_h),a_h) - \mathbb{E}_{\substack{s_{h+1}\sim\mathbb{T}_h(\cdot\mid s_h,a_h),\\ o_{h+1}\sim\mathbb{O}_{h+1}(\cdot\mid s_{h+1})}}\left[r_h(s_h,a_h) + \widetilde{V}_{h+1}^{\pi}(z_{h+1},s_{h+1})\right]\right] \\
&\geq \sum_{h\in[H]} \mathbb{E}_{\overline{\tau}_h\sim\pi^*}\left[\left\langle \widetilde{Q}_h^{\pi^t}((z_h,s_h),\cdot), \pi_h^t(\cdot \mid z_h,s_h) - \pi_h^*(\cdot \mid z_h,s_h)\right\rangle\right],
\end{aligned}
$$

where in the inequality above we used Lemma H.2. Since our policy does not depend on the realized latent state s_h,

$$
\begin{aligned}
&\widetilde{V}_1^{\pi^t}(s_1) - V_1^{\pi^*}(s_1) \\
&\geq \sum_{h\in[H]} \mathbb{E}_{\overline{\tau}_h\sim\pi^*}\left[\left\langle \widetilde{Q}_h^{\pi^t}((z_h,s_h),\cdot), \pi_h^t(\cdot \mid z_h,s_h) - \pi_h^*(\cdot \mid z_h,s_h)\right\rangle\right] \\
&= \sum_{h\in[H]} \mathbb{E}_{\tau_h\sim\pi^*}\left[\left\langle \mathbb{E}_{s_h\sim\boldsymbol{b}(\tau_h)}\left[\widetilde{Q}_h^{\pi^t}((z_h,s_h),\cdot)\right], \pi_h^t(\cdot \mid z_h) - \pi_h^*(\cdot \mid z_h)\right\rangle\right] \\
&= \sum_{h\in[H]} \mathbb{E}_{\tau_h\sim\pi^*}\left[\left\langle \mathbb{E}_{s_h\sim\boldsymbol{b}^{\text{apx}}(z_h)}\left[\widetilde{Q}_h^{\pi^t}((z_h,s_h),\cdot)\right], \pi_h^t(\cdot \mid z_h) - \pi_h^*(\cdot \mid z_h)\right\rangle\right] \\
&\quad + \sum_{h\in[H]} \mathbb{E}_{\tau_h\sim\pi^*}\left[\left\langle \mathbb{E}_{s_h\sim\boldsymbol{b}(\tau_h)}\left[\widetilde{Q}_h^{\pi^t}((z_h,s_h),\cdot)\right] - \mathbb{E}_{s_h\sim\boldsymbol{b}^{\text{apx}}(z_h)}\left[\widetilde{Q}_h^{\pi^t}((z_h,s_h),\cdot)\right], \pi_h^t(\cdot \mid z_h) - \pi_h^*(\cdot \mid z_h)\right\rangle\right] \\
&\geq \sum_{h\in[H]} \mathbb{E}_{\tau_h\sim\pi^*}\left[\left\langle \mathbb{E}_{s_h\sim\boldsymbol{b}^{\text{apx}}(z_h)}\left[\widetilde{Q}_h^{\pi^t}((z_h,s_h),\cdot)\right], \pi_h^t(\cdot \mid z_h) - \pi_h^*(\cdot \mid z_h)\right\rangle\right] \\
&\quad - \sum_{h\in[H]} \mathbb{E}_{\tau_h\sim\pi^*}\left[\left\|\mathbb{E}_{s_h\sim\boldsymbol{b}(\tau_h)}\left[\widetilde{Q}_h^{\pi^t}((z_h,s_h),\cdot)\right] - \mathbb{E}_{s_h\sim\boldsymbol{b}^{\text{apx}}(z_h)}\left[\widetilde{Q}_h^{\pi^t}((z_h,s_h),\cdot)\right]\right\|_1\right] \\
&\geq \sum_{h\in[H]} \mathbb{E}_{\tau_h\sim\pi^*}\left[\left\langle \mathbb{E}_{s_h\sim\boldsymbol{b}^{\text{apx}}(z_h)}\left[\widetilde{Q}_h^{\pi^t}((z_h,s_h),\cdot)\right], \pi_h^t(\cdot \mid z_h) - \pi_h^*(\cdot \mid z_h)\right\rangle\right] \\
&\quad - 2\cdot H\cdot \sum_{h\in[H]} \mathbb{E}_{\tau_h\sim\pi^*}\left[d_{TV}(\boldsymbol{b}_h(\tau_h), \boldsymbol{b}_h^{\text{apx}}(z_h))\right].
\end{aligned}
$$

The last inequality follows by assumption Lemma H.2. By averaging we get,

$$
\begin{aligned}
\frac{1}{T}\sum_{t\in[T]} \widetilde{V}_1^{\pi^t}(s_1) \geq & V_1^{\pi^*}(s_1) + \frac{1}{T}\sum_{h\in[H]} \mathbb{E}_{\tau_h\sim\pi^*}\left[\sum_{t\in[T]}\left\langle \mathbb{E}_{s_h\sim\boldsymbol{b}^{\text{apx}}(z_h)}\left[\widetilde{Q}_h^{\pi^t}((z_h,s_h),\cdot)\right], \pi_h^t(\cdot \mid z_h) - \pi_h^*(\cdot \mid z_h)\right\rangle\right] \\
& - 2\cdot H\cdot \sum_{h\in[H]} \mathbb{E}_{\tau_h\sim\pi^*}\left[d_{TV}(\boldsymbol{b}_h(\tau_h), \boldsymbol{b}_h^{\text{apx}}(z_h))\right]. \\
\geq & V_1^{\pi^*}(s_1) + \frac{H}{T}\max_{h\in[H]} \mathbb{E}_{\tau_h\sim\pi^*}\left[\sum_{t\in[T]}\left\langle \mathbb{E}_{s_h\sim\boldsymbol{b}^{\text{apx}}(z_h)}\left[\widetilde{Q}_h^{\pi^t}((z_h,s_h),\cdot)\right], \pi_h^t(\cdot \mid z_h) - \pi_h^*(\cdot \mid z_h)\right\rangle\right] \\
& - 2\cdot H^2\cdot \max_{h\in[H]} \mathbb{E}_{\tau_h\sim\pi^*}\left[d_{TV}(\boldsymbol{b}_h(\tau_h), \boldsymbol{b}_h^{\text{apx}}(z_h))\right]. \\
\geq & V_1^{\pi^*}(s_1) - \frac{2H\sqrt{H\log(|\mathcal{A}|)}}{\sqrt{T}} - 2\cdot H^2\cdot \max_{h\in[H]} \mathbb{E}_{\tau_h\sim\pi^*}\left[d_{TV}(\boldsymbol{b}_h(\tau_h), \boldsymbol{b}_h^{\text{apx}}(z_h))\right],
\end{aligned}
$$

[3]Since π^* is deterministic, w.l.o.g. we define by $a^*_{z_h}$ the action that the best-in-class policy selects at truncated memory z_h.

where the last inequality follows since for fixed $h \in [H]$ and $z_h \in \mathcal{Z}_h$, the agent updates her policy on memory z_h according to MWU on feedback $\left\{\mathbb{E}_{s_h \sim \boldsymbol{b}^{\text{apx}}(z_h)}\left[\widetilde{Q}_h^{\pi^t}((z_h, s_h), a)\right]\right\}_{a \in \mathcal{A}}$, and thus, the accumulate regret is bounded by (Section 4.3 in [8]):

$$\begin{aligned}\sum_{t \in [T]} \left\langle \mathbb{E}_{s_h \sim \boldsymbol{b}^{\text{apx}}(z_h)}\left[\widetilde{Q}_h^{\pi^t}((z_h, s_h)\right], \pi_h^*(\cdot \mid z_h) - \pi_h^t(\cdot \mid z_h)\right\rangle \leq & \frac{\log(|\mathcal{A}|)}{\eta} + \eta \cdot T \cdot \left\| Q_h^{\pi^t}((z_h, s_h), \cdot)\right\|_{+\infty} \\ \leq & \frac{\log(|\mathcal{A}|)}{\eta} + \eta \cdot T \cdot H \\ = & 2\sqrt{T \cdot H \log(|\mathcal{A}|)}.\end{aligned}$$

The proof follows by combining Equation (H.1), and the inequality above. Finally, to achieve the near optimality in the class of Π^L, we bound the optimistic estimation using Equation (H.2) in Lemma H.2, and its global optimality under γ-observability is a direct consequence of [22]. ■

Lemma H.2 (Optimistic Q-function - adapted from [40])**.** Given a policy $\pi \in \Pi^L$, and a parameter $M \in \mathbb{N}$, let $\{\widetilde{Q}_h^\pi : \mathcal{Z}_h \times \mathcal{A} \to [0, H]\}_{h \in [H]}$ be the output of Algorithm 3. Then with probability at least $1 - \delta$:

$$H - h + 1 \geq \widetilde{Q}_h^\pi((z_h, s_h), a_h) \geq \mathbb{E}_{\substack{s_{h+1} \sim \mathbb{T}_h(\cdot \mid s_h, a_h), \\ o_{h+1} \sim \mathbb{O}_{h+1}(\cdot \mid s_{h+1})}}\left[r_h(s_h, a_h) + \widetilde{V}_{h+1}^\pi(z_{h+1}, s_{h+1})\right], \forall z_h \in \mathcal{Z}_h, s_h \in \mathcal{S}, a_h \in \mathcal{A},$$

$$\widetilde{V}_1^\pi(s_1) - V^\pi(s_1) \leq O\left(H^2 \cdot \sqrt{\frac{\max(|\mathcal{O}|, |\mathcal{S}|) \cdot |\mathcal{S}| \cdot |\mathcal{A}|}{M} \cdot \log\left(\frac{|\mathcal{S}| \cdot |\mathcal{A}|}{\delta}\right) \log\left(\frac{M \cdot |\mathcal{S}| \cdot |\mathcal{A}| \cdot H}{\delta}\right)}\right), \tag{H.2}$$

where $\widetilde{V}_h^\pi(z_h, s_h) = \mathbb{E}_{a_h \sim \pi(z_h)}[\widetilde{Q}_h^\pi(z_h, a_h)]$. Moreover, Algorithm 3 needs a total of $H \cdot M$ episodes from POMDP $\mathcal{P}$ and runs in time $\text{POLY}(H, M, |\mathcal{A}|^L, |\mathcal{O}|^L)$.

Proof. For each step $h \in [H]$, collect M trajectories with states using policy π on POMDP $\mathcal{P}$ and let $D_h = \{\overline{\tau}^{(i)}\}_{i \in [M]}$ be those collected trajectories. Define the empirical transition, observation and reward distribution as follows:

$$\begin{aligned}N_h(s_h, a_h, s_{h+1}) = & |\overline{\tau} = (s_1', o_1', a_1', r_1' \ldots, s_h', o_h', a_h', r_h') \in D_h : (s_h, a_h, s_{h+1}) = (s_h', a_h', s_{h+1}')|, \\ N_h(s_h, a_h) = & \sum_{s_{h+1} \in \mathcal{S}} N_h(s_h, a_h, s_{h+1}), \\ N_h(s_h) = & \sum_{a_h \in \mathcal{A}} N_h(s_h, a_h), \\ N_h(s_h, o_h) = & |\tau = (s_1', o_1', a_1', r_1' \ldots, s_h', o_h', a_h', r_h') \in D_h : (s_h, o_h) = (s_h', o_h')|, \\ \widehat{\mathbb{T}}_h(s_{h+1} \mid s_h, a_h) = & \frac{N_h(s_h, a_h, s_{h+1})}{N_h(s_h, a_h)}, \\ \widehat{\mathbb{O}}_h(o_h \mid s_h) = & \frac{N_h(s_h, o_h)}{N_h(s_h)}.\end{aligned}$$

Set $\delta_1 = \frac{\delta}{2 \cdot |\mathcal{S}| \cdot (|\mathcal{A}|+1)}$. By [10], there exists a constant $C > 0$ such that for each step $h \in [H]$, state $s \in \mathcal{S}$ and action $a \in \mathcal{A}$ with probability at least $1 - \delta_1$:

$$\begin{aligned}\|\mathbb{T}_h(\cdot \mid s_h, a_h) - \widehat{\mathbb{T}}_h(\cdot \mid s_h, a_h)\|_1 \leq & \min\left(2, C \cdot \sqrt{\frac{|\mathcal{S}| \log(1/\delta_1)}{\max(N_h(s_h, a_h), 1)}}\right), \\ \|\mathbb{O}_h(\cdot \mid s_h) - \widehat{\mathbb{O}}_h(\cdot \mid s_h)\|_1 \leq & \min\left(2, C \cdot \sqrt{\frac{|\mathcal{O}| \log(1/\delta_1)}{\max(N_h(s_h), 1)}}\right).\end{aligned}$$

For the rest of the proof we condition on this event. By union bound, this happen with probability at least $1 - \frac{\delta}{2}$. We define the optimistic Q function recursively as follows for a memory-state pair

$(z_h, s_h) \in \mathcal{Z}_h \times \mathcal{S}$:

$$\begin{aligned}
\widetilde{Q}^{\pi}_{H+1}(z_{H+1}, s_{H+1}) =& 0, \qquad \forall z_{H+1} \in \mathcal{Z}_{H+1}, s_{H+1} \in \mathcal{S} \\
\widetilde{Q}^{\pi}_{h}((z_h, s_h), a_h) =& \min \Bigg(H - h + 1, \mathbb{E}_{\substack{s_{h+1} \sim \widehat{\mathbb{T}}_h(s_h, a_h), \\ o_{h+1} \sim \widehat{\mathbb{O}}_{h+1}(s_{h+1})}} [\widetilde{V}^{\pi}_{h+1}(z_{h+1}, s_{h+1})] \\
& + r(s_h, a_h) + H \cdot \min \left(2, C \cdot \sqrt{\frac{|\mathcal{S}| \log(1/\delta_1)}{\max(N_h(s_h, a_h), 1)}} \right) \\
& + \mathbb{E}_{s_{h+1} \sim \widehat{\mathbb{T}}_h(s_h, a_h)} H \cdot \min \left(2, C \cdot \sqrt{\frac{|\mathcal{O}| \log(1/\delta_1)}{\max(N_{h+1}(s_{h+1}), 1)}} \right) \Bigg),
\end{aligned}$$

where $\widetilde{V}^{\pi}_h(z_h, s_h) = \mathbb{E}_{a_h \sim \pi(z_h)}[\widetilde{Q}^{\pi}_h(z_h, a_h)]$. Hence the time complexity of our algorithm is $\text{POLY}(H, M, |\mathcal{A}|^L, |\mathcal{O}|^L)$. To prove the first condition we fix step $h \in [H], z_h \in \mathcal{Z}_h, a_h \in \mathcal{A}$ and state $s_h \in \mathcal{S}$ and take condition whether $\widetilde{Q}^{\pi}_h(z_h, s_h) = H - h + 1$. In this case, since by assumption on the POMDP $\mathcal{P}$, $r_h(s_h, a_h) \leq 1$, and by definition of $\widetilde{Q}^{\pi}_{h+1}(\cdot, \cdot) \leq H - h$ we have:

$$\widetilde{Q}^{\pi}_h((z_h, s_h), a_h) = 1 + H - h \geq \mathbb{E}_{\substack{s_{h+1} \sim \mathbb{T}_h(s_h, a_h), \\ o_{h+1} \sim \mathbb{O}_{h+1}(s_{h+1})}} [r_h(s_h, a_h) + \widetilde{V}^{\pi}_{h+1}(z_{h+1}, s_{h+1})].$$

If $\widetilde{Q}^{\pi}_h(z_h, s_h) \neq H - h + 1$, observe that:

$$\begin{aligned}
\widetilde{Q}^{\pi}_h((z_h, s_h), a_h) =& \mathbb{E}_{\substack{s_{h+1} \sim \widehat{\mathbb{T}}_h(s_h, a_h), \\ o_{h+1} \sim \widehat{\mathbb{O}}_{h+1}(s_{h+1})}} [\widetilde{V}^{\pi}_{h+1}(z_{h+1}, s_{h+1})] \\
& + r(s_h, a_h) + H \cdot \min \left(2, C \cdot \sqrt{\frac{|\mathcal{S}| \log(1/\delta_1)}{\max(N_h(s_h, a_h), 1)}} \right) \\
& + \mathbb{E}_{s_{h+1} \sim \widehat{\mathbb{T}}_h(s_h, a_h)} H \cdot \min \left(2, C \cdot \sqrt{\frac{|\mathcal{O}| \log(1/\delta_1)}{\max(N_{h+1}(s_{h+1}), 1)}} \right) \\
\geq& \mathbb{E}_{\substack{s_{h+1} \sim \mathbb{T}_h(s_h, a_h), \\ o_{h+1} \sim \mathbb{O}_{h+1}(s_{h+1})}} [r_h(s_h, a_h) + \widetilde{V}^{\pi}_{h+1}(z_{h+1}, s_{h+1})],
\end{aligned}$$

and hence, $\{\widetilde{Q}^{\pi}_h\}_{h \in [H]}$ satisfies the first condition. Moreover, it is true that:

$$\begin{aligned}
\widetilde{Q}^{\pi}_h((z_h, s_h), a_h) \leq& \mathbb{E}_{\substack{s_{h+1} \sim \mathbb{T}_h(s_h, a_h), \\ o_{h+1} \sim \mathbb{O}_{h+1}(s_{h+1})}} [r_h(s_h, a_h) + \widetilde{V}^{\pi}_{h+1}(z_{h+1}, s_{h+1})] \\
& + 2H \cdot \min \left(2, C \cdot \sqrt{\frac{|\mathcal{S}| \log(1/\delta_1)}{\max(N_h(s_h, a_h), 1)}} \right) \\
& + 2 \cdot \mathbb{E}_{s_{h+1} \sim \widehat{\mathbb{T}}_h(s_h, a_h)} H \cdot \min \left(2, C \cdot \sqrt{\frac{|\mathcal{O}| \log(1/\delta_1)}{\max(N_{h+1}(s_{h+1}), 1)}} \right) \\
\leq& \mathbb{E}_{\substack{s_{h+1} \sim \mathbb{T}_h(s_h, a_h), \\ o_{h+1} \sim \mathbb{O}_{h+1}(s_{h+1})}} [r_h(s_h, a_h) + \widetilde{V}^{\pi}_{h+1}(z_{h+1}, s_{h+1})] \\
& + 6H \cdot \min \left(2, C \cdot \sqrt{\frac{|\mathcal{S}| \log(1/\delta_1)}{\max(N_h(s_h, a_h), 1)}} \right) \\
& + 2 \cdot \mathbb{E}_{s_{h+1} \sim \mathbb{T}_h(s_h, a_h)} H \cdot \min \left(2, C \cdot \sqrt{\frac{|\mathcal{O}| \log(1/\delta_1)}{\max(N_{h+1}(s_{h+1}), 1)}} \right).
\end{aligned}$$

Thus it holds that:

$$\begin{aligned}\widetilde{V}_h^\pi(z_h,s_h)-V_h^\pi(z_h,s_h)\leq&\mathbb{E}_{\substack{a_h\sim\pi(z_h),\\ s_{h+1}\sim\mathbb{T}_h(s_h,a_h),\\ o_{h+1}\sim\mathbb{O}_{h+1}(s_{h+1})}}[\widetilde{V}_{h+1}^\pi(z_{h+1},s_{h+1})-V_{h+1}^\pi(z_{h+1},s_{h+1})]\\&+6\cdot C\cdot H\cdot\mathbb{E}_{a_h\sim\pi(z_h)}\sqrt{\frac{|\mathcal{S}|\log(1/\delta_1)}{\max(N_h(s_h,a_h),1)}}\\&+2\cdot C\cdot H\cdot\mathbb{E}_{\substack{a_h\sim\pi(z_h),\\ s_{h+1}\sim\mathbb{T}_h(s_h,a_h)}}\sqrt{\frac{|\mathcal{O}|\log(1/\delta_1)}{\max(N_{h+1}(s_{h+1}),1)}}.\end{aligned}$$

Thus we conclude that

$$\begin{aligned}\widetilde{V}_1^\pi(s_1)-V_1^\pi(s_1)\leq&\mathbb{E}_{\tau=(s_1,a_1,\ldots,s_{H+1})\sim\pi}\left[\sum_{h\in[H]}8\cdot H\cdot C\cdot\sqrt{\frac{\max(|\mathcal{S}|,|\mathcal{O}|)\log(1/\delta_1)}{\max(N_h(s_h,a_h),1)}}\right]\\=&8\cdot H\sqrt{\max(|\mathcal{O}|,|\mathcal{S}|)\cdot\log(1/\delta_1)}\cdot C\cdot\sum_{h\in[H]}\mathbb{E}_{\tau=(s_1,a_1,\ldots,s_{H+1})\sim\pi}\left[\sqrt{\frac{1}{\max(N_h(s_h,a_h),1)}}\right].\end{aligned}$$

To finish the proof, we make use of the following lemma.

Lemma H.3 (Lemma 6 in [40])**.** For each step $h\in[H]$, and state-action pair $s_h,a_h\in\mathcal{S}\times\mathcal{A}$, with probability at least $1-\delta_2$:

$$\sqrt{\frac{1}{\max(N_h(s_h,a_h),1)}}=O\left(\sqrt{\frac{|\mathcal{S}|\cdot|\mathcal{A}|\log(M/\delta_2)}{M}}\right).$$

By setting $\delta_2=\frac{\delta}{2\cdot|\mathcal{S}|\cdot|\mathcal{A}|\cdot H}$, by taking union bound we have that with probability at least $1-\delta$ we conclude that:

$$\begin{aligned}\widetilde{V}_1^\pi(s_1)-V_1^\pi(s_1)=&8\sqrt{\max(|\mathcal{O}|,|\mathcal{S}|)\cdot\log(1/\delta_1)}\cdot C\cdot\sum_{h\in[H]}\mathbb{E}_{\tau=(s_1,a_1,\ldots,s_{H+1})\sim\pi}\left[\sqrt{\frac{1}{\max(N_h(s_h,a_h),1)}}\right]\\\leq&O\left(H^2\cdot\sqrt{\frac{\max(|\mathcal{O}|,|\mathcal{S}|)\cdot|\mathcal{S}|\cdot|\mathcal{A}|}{M}}\cdot\log\left(\frac{|\mathcal{S}|\cdot|\mathcal{A}|}{\delta}\right)\log\left(\frac{M\cdot|\mathcal{S}|\cdot|\mathcal{A}|\cdot H}{\delta}\right)\right).\end{aligned}$$

□

The proof of Theorem 5.2 follows by combining Theorem H.4 and Theorem H.5 below. Theorem H.4 proves that we can approximately learn a POMDP model $\mathcal{P}$ computational and sample efficiently, thanks to the privileged information.

Theorem H.4. Fix any $\epsilon,\delta\in(0,1)$. Algorithm 4 can learn the approximate POMDP model with transition $\widehat{\mathbb{T}}_{1:H}$ and emission $\widehat{\mathbb{O}}_{1:H}$ such that with probability $1-\delta$, for any policy $\pi\in\Pi^{\text{gen}}$ and $h\in[H]$

$$\mathbb{E}_\pi^{\mathcal{P}}\left[\|\mathbb{T}_h(\cdot\,|\,s_h,a_h)-\widehat{\mathbb{T}}_h(\cdot\,|\,s_h,a_h)\|_1+\|\mathbb{O}_h(\cdot\,|\,s_h)-\widehat{\mathbb{O}}_h(\cdot\,|\,s_h)\|_1\right]\leq\epsilon,$$

using $\text{POLY}(S,A,H,O,\frac{1}{\epsilon},\log(\frac{1}{\delta}))$ episodes in time $\text{POLY}(S,A,H,O,\frac{1}{\epsilon},\log(\frac{1}{\delta}))$.

Proof. Note that by Lemma H.8, it suffices to consider only $\pi\in\Pi_{\mathcal{S}}$ (by considering $r_h(s_h,a_h):=\|\mathbb{T}_h(\cdot\,|\,s_h,a_h)-\widehat{\mathbb{T}}_h(\cdot\,|\,s_h,a_h)\|_1+\|\mathbb{O}_h(\cdot\,|\,s_h)-\widehat{\mathbb{O}}_h(\cdot\,|\,s_h)\|_1$). For each $h\in[H]$ and $s_h\in\mathcal{S}$, we define

$$p_h(s_h)=\max_{\pi\in\Pi_{\mathcal{S}}}d_h^\pi(s_h).$$

Fix $\epsilon_1 > 0$, we define $\mathcal{U}(h, \epsilon_1) = \{s_h \in \mathcal{S} \mid p_h(s_h) \geq \epsilon_1\}$. By [32], one can learn the policy $\Psi(h, s_h)$ with sample complexity $\widetilde{\mathcal{O}}(\frac{S^2AH^4}{\epsilon_1})$ such that $d_h^{\Psi(h,s_h)}(s_h) \geq \frac{p_h(s_h)}{2}$ for each $s_h \in \mathcal{U}(h, \epsilon_1)$ with probability $1 - \delta_1$. Now we assume this event holds for any $h \in [H]$ and $s_h \in \mathcal{U}(h, \epsilon_1)$. For each $s_h \in \mathcal{S}$ and $a_h \in \mathcal{A}$, we have executed the policy $\Psi(h, s_h)$ followed by an action $a_h \in \mathcal{A}$ for N episodes and denote the number of episodes that s_h and a_h are visited as $N_h(s_h, a_h)$. Then with probability $1 - e^{N\epsilon_1/8}$, $N_h(s_h, a_h) \geq \frac{Np_h(s_h)}{2}$ by Chernoff bound. Now conditioned on this event, we are ready to evaluate the following

$$
\begin{aligned}
&\mathbb{E}^{\mathcal{P}}_{\Psi(h,s_h)} \|\mathbb{T}_h(\cdot \mid s_h, a_h) - \widehat{\mathbb{T}}_h(\cdot \mid s_h, a_h)\|_1 \\
&= \sum_{s_h, a_h} d_h^{\Psi(h,s_h)}(s_h)\Psi(h, s_h)_h(a_h \mid s_h)\|\mathbb{T}_h(\cdot \mid s_h, a_h) - \widehat{\mathbb{T}}_h(\cdot \mid s_h, a_h)\|_1 \\
&\leq 2S\epsilon_1 + \sum_{s_h \in \mathcal{U}(h,\epsilon_1), a_h} d_h^{\Psi(h,s_h)}(s_h)\Psi(h, s_h)_h(a_h \mid s_h)\sqrt{\frac{S\log(1/\delta_2)}{Np_h(s_h)}} \\
&\leq 2S\epsilon_1 + \sum_{s_h} \sqrt{d_h^{\Psi(h,s_h)}(s_h)}\sqrt{\frac{S\log(1/\delta_2)}{N}} \\
&\leq 2S\epsilon_1 + S\sqrt{\frac{\log(1/\delta_2)}{N}},
\end{aligned}
$$

where the first inequality follows by [10] with probability at least $1 - \delta_2$. Similarly,

$$
\begin{aligned}
&\mathbb{E}^{\mathcal{P}}_{\Psi(h,s_h)} \|\mathbb{O}_h(\cdot \mid s_h) - \widehat{\mathbb{O}}_h(\cdot \mid s_h)\|_1 \\
&= \sum_{s_h} d_h^{\Psi(h,s_h)}(s_h)\|\mathbb{O}_h(\cdot \mid s_h) - \widehat{\mathbb{O}}_h(\cdot \mid s_h)\|_1 \\
&\leq S\epsilon_1 + \sum_{s_h \in \mathcal{U}(h,\epsilon_1)} d_h^{\Psi(h,s_h)}(s_h)\sqrt{\frac{O\log(1/\delta_2)}{Np_h(s_h)}} \\
&\leq S\epsilon_1 + \sum_{s_h \in \mathcal{U}(h,\epsilon_1)} \sqrt{d_h^{\Psi(h,s_h)}(s_h)}\sqrt{\frac{O\log(1/\delta_2)}{N}} \\
&\leq S\epsilon_1 + \sqrt{\frac{SO\log(1/\delta_2)}{N}}.
\end{aligned}
$$

Therefore, by a union bound, the all high probability events hold with probability

$$1 - SH\delta_1 - SHAe^{N\epsilon_1/8} - SAH\delta_2,$$

where similarly the first inequality follows by [10] with probability at least $1 - \delta_2$. Therefore, we can choose $N = \widetilde{\mathcal{O}}(\frac{S^2+SO}{\epsilon^2})$ and $\epsilon_1 = \mathcal{O}(\frac{\epsilon}{S})$, leading to the total sample complexity

$$SHA(N + \widetilde{\mathcal{O}}(\frac{S^3AH^4}{\epsilon})) = \widetilde{\mathcal{O}}(\frac{S^2AHO + S^3AH}{\epsilon^2} + \frac{S^4A^2H^5}{\epsilon}). \tag{H.3}$$

□

Now with such a learned model in a reward-free way, we are ready to present our main result for approximate belief learning.

Theorem H.5. Consider a γ-observable POMDP $\mathcal{P}$, an $\epsilon > 0$ and let $\widehat{\mathcal{P}}$ be the outcome of Algorithm 4 such that for any $\pi \in \Pi^{\text{gen}}$:

$$\mathbb{E}^{\mathcal{P}}_{\pi}\left[\|\mathbb{T}_h(\cdot \mid s_h, a_h) - \widehat{\mathbb{T}}_h(\cdot \mid s_h, a_h)\|_1 + \|\mathbb{O}_h(\cdot \mid s_h) - \widehat{\mathbb{O}}_h(\cdot \mid s_h)\|_1\right] \leq \frac{\epsilon}{3 \cdot H}.$$

Then we can construct in time $\text{POLY}(H, |\mathcal{A}|, |\mathcal{S}|, |\mathcal{O}|, \frac{1}{\epsilon}, \log\frac{1}{\delta})$ a belief $\{\boldsymbol{b}_h^{\text{apx}} : \mathcal{Z}_h \to \Delta(\mathcal{S})\}_{h\in[H]}$ with no further samples. In addition if $N \geq \mathcal{O}(\frac{O\log(SH/\delta)}{\gamma^2\epsilon_1})$ and $L \geq \widetilde{\Omega}(\gamma^{-4}\log(S/\epsilon))$, then for any $\pi \in \Pi^{\text{gen}}$ and $h \in [H]$:

$$\mathbb{E}^{\mathcal{P}}_{\pi}\|\boldsymbol{b}_h(\tau_h) - \boldsymbol{b}_h^{\text{apx}}(z_h)\|_1 \leq \epsilon.$$

Proof. We firstly consider the following simple while important fact: for the estimated emission $\widehat{\mathbb{O}}_h$, its observability can be evaluated as

$$\|\widehat{\mathbb{O}}_h^\top(b-b')\|_1 \geq \|\mathbb{O}_h^\top(b-b')\|_1 - \|(\mathbb{O}_h^\top - \widehat{\mathbb{O}}_h^\top)(b-b')\|_1 \geq (\gamma - \|\widehat{\mathbb{O}}_h - \mathbb{O}_h\|_\infty)\|b-b'\|_1,$$

for any $b, b' \in \Delta(\mathcal{S})$ and $\|\widehat{\mathbb{O}}_h - \mathbb{O}_h\|_\infty := \max_{s_h \in \mathcal{S}} \|\mathbb{O}_h(\cdot \,|\, s_h) - \widehat{\mathbb{O}}_h(\cdot \,|\, s_h)\|_1$. Therefore, if one can ensure that the emission at any state s_h is learned accurately in the sense that $\|\mathbb{O}_h(\cdot \,|\, s_h) - \widehat{\mathbb{O}}_h(\cdot \,|\, s_h)\|_1 \leq \frac{\gamma}{2}$, we can conclude that $\widehat{\mathbb{O}}_h$ is also $\gamma/2$-observable. However, the key challenge here is that there could exist some states s_h at step h that can only be visited with a low probability no matter what exploration policy is used. Therefore, emissions at such states may not be learned accurately.

To circumvent this issue, our key technique here is to *redirect* the transition probability into states that cannot be explored sufficiently to a new state s^{exit}, the emission associated with which gives a new observation o^{exit} deterministically. For ease of our presentation, we assume that we know that there already exists such a dummy pair of $(s^{\text{exit}}, o^{\text{exit}})$ in the POMDP model $\mathcal{P}$ we are dealing with. Formally, we assume that $s^{\text{exit}} \in \mathcal{S}$ is not reachable at each step $h \in [H]$, and the system state keeps staying at s^{exit}. Meanwhile, we assume $o^{\text{exit}} \in \mathcal{O}$ satisfies that $\mathbb{O}_h(o^{\text{exit}} \,|\, s^{\text{exit}}) = 1$, while $\mathbb{O}_h(o^{\text{exit}} \,|\, s) = 0$ for any $s \neq s^{\text{exit}}$ and $h \in [H]$.

To see why we can assume we know such an $(s^{\text{exit}}, o^{\text{exit}})$, we notice that if $\mathcal{P}$ does not contain such a pair (or such a pair does exist, but we don't know which state-observation it is), we can construct a new $\mathcal{P}'$ by manually adding such $(s^{\text{exit}}, o^{\text{exit}})$ to $\mathcal{P}$ by letting the transition probability at any other states to s^{exit} to be 0, and also manually adding the dummy observation o^{exit} to satisfy the requirement above. By such a construction, any ϵ-optimal policy of $\mathcal{P}'$ is an ϵ-optimal policy of $\mathcal{P}$ since s^{exit} is not reachable. More importantly, simulating $\mathcal{P}'$ is exactly equivalent to simulating the original $\mathcal{P}$ again since s^{exit} is not reachable. Finally, we point out that if the original problem $\mathcal{P}$ is γ-observable, we have $\mathcal{P}'$ is also γ-observable. Then we only need to deal with the new model $\mathcal{P}'$ that satisfies our requirement.

Now, for any $\epsilon_1 > 0$, we define

$$\mathcal{S}_h^{\text{low}} := \left\{ s_h \in \mathcal{S} \,|\, \frac{N_h(s_h)}{N} < \frac{\epsilon_1}{2} \right\}, \quad \mathcal{S}_h^{\text{high}} := \mathcal{S} \setminus \mathcal{S}_h^{\text{low}}.$$

With Chernoff bound, with probability at least $1 - Se^{-N\epsilon_1/8}$, it holds that

$$\mathcal{S}_h^{\text{low}} \subseteq \{s_h \in \mathcal{S} \,|\, p_h(s_h) < \epsilon_1\},$$

where $p_h(s_h) := \max_{\pi \in \Pi_{\mathcal{S}}} d_h^\pi(s_h)$. To see the reason, we notice that for any s_h such that $p_h(s_h) \geq \epsilon_1$, with probability $1 - e^{-N\epsilon_1/8}$, it holds that $\frac{N_h(s_h)}{N} \geq \frac{\epsilon_1}{2}$. Therefore, the last step is by taking a union bound for all s_h. Now with $\mathcal{S}_h^{\text{low}}$ defined, we are ready to construct a truncated POMDP $\mathcal{P}^{\text{trunc}}$ such that for each $h \in [H]$, we define the transition as

$$\mathbb{T}_h^{\text{trunc}}(s_{h+1} \,|\, s_h, a_h) := \mathbb{T}_h(s_{h+1} \,|\, s_h, a_h), \forall s_h \in \mathcal{S}_h^{\text{high}}, s_{h+1} \in \mathcal{S}_{h+1}^{\text{high}}, a_h \in \mathcal{A}.$$

Meanwhile, we define

$$\mathbb{T}_h^{\text{trunc}}(s^{\text{exit}} \,|\, s_h, a_h) := \sum_{s_{h+1} \in \mathcal{S}_{h+1}^{\text{low}}} \mathbb{T}_h(s_{h+1} \,|\, s_h, a_h), \forall s_h \in \mathcal{S}_h^{\text{high}}, a_h \in \mathcal{A}.$$

Recall that once the state becomes s^{exit}, future states will always be s^{exit} no matter what actions are taken. Finally, we define the transition for $s_h \in \mathcal{S}_h^{\text{low}}$ as,

$$\mathbb{T}_h^{\text{trunc}}(\cdot \,|\, s_h, \cdot) := \mathbb{T}_h(\cdot \,|\, s_h, \cdot).$$

For emission, we simply define

$$\mathbb{O}_h^{\text{trunc}}(o_h \,|\, s_h) := \mathbb{O}_h(o_h \,|\, s_h), \forall h \in [H], s_h \in \mathcal{S}, o_h \in \mathcal{O}_h.$$

Finally, we define the rewards for $\mathcal{P}^{\text{trunc}}$ arbitrarily. Now we examine the total variation distance of trajectory distribution in $\mathcal{P}$ and $\mathcal{P}^{\text{trunc}}$. For any policy $\pi \in \Pi^{\text{gen}}$, it is easy to see that

$$\mathbb{P}^{\pi,\mathcal{P}}(\overline{\tau}_H) = \mathbb{P}^{\pi,\mathcal{P}^{\text{trunc}}}(\overline{\tau}_H),$$

for any $\overline{\tau}_H \in \overline{\mathcal{T}}_H^{\text{high}} := \{(s_{1:H}, o_{1:H}, a_{1:H}) \mid s_h \in \mathcal{S}_h^{\text{high}}, \forall h \in [H]\}$. Meanwhile, it holds by a union bound that

$$\mathbb{P}^{\pi,\mathcal{P}}(\overline{\tau}_H \notin \overline{\mathcal{T}}_H^{\text{high}}) \le \sum_{h\in[H]} \mathbb{P}^{\pi,\mathcal{P}}(s_h \in \mathcal{S}_h^{\text{low}}) \le HS\epsilon_1.$$

Therefore, the total variation distance of trajectory distribution in $\mathcal{P}$ and $\mathcal{P}^{\text{trunc}}$ can be bounded by

$$\sum_{\overline{\tau}_H} |\mathbb{P}^{\pi,\mathcal{P}}(\overline{\tau}_H) - \mathbb{P}^{\pi,\mathcal{P}^{\text{trunc}}}(\overline{\tau}_H)| \le 2HS\epsilon_1.$$

On the other hand, by Lemma H.7, we have

$$\begin{aligned}
\mathbb{E}_\pi^{\mathcal{P}}[\|\boldsymbol{b}_h(\tau_h) - \boldsymbol{b}_h^{\text{trunc}}(\tau_h)\|_1] &\le \mathbb{E}_\pi^{\mathcal{P}} \sum_{t\in[h]} \|\mathbb{T}_t(\cdot \mid s_t, a_t) - \mathbb{T}_t^{\text{trunc}}(\cdot \mid s_t, a_t)\|_1 + \|\mathbb{O}_t(\cdot \mid s_t) - \mathbb{O}_t^{\text{trunc}}(\cdot \mid s_t)\|_1 \\
&\le 4\mathbb{P}^{\pi,\mathcal{P}}(\exists t \in [h] : s_t \notin \mathcal{S}_t^{\text{high}}) \\
&\le 4HS\epsilon_1.
\end{aligned}$$

With such an intermediate quantity $\mathcal{P}^{\text{trunc}}$, we define its approximate version $\widehat{\mathcal{P}}^{\text{trunc}}$ as follows:

$$\widehat{\mathbb{T}}_h^{\text{trunc}}(s_{h+1} \mid s_h, a_h) := \widehat{\mathbb{T}}_h(s_{h+1} \mid s_h, a_h) \quad \forall h \in [H], s_h \in \mathcal{S}_h^{\text{high}}, s_{h+1} \in \mathcal{S}_{h+1}^{\text{high}}, a_h \in \mathcal{A},$$

and

$$\widehat{\mathbb{T}}_h^{\text{trunc}}(s^{\text{exit}} \mid s_h, a_h) := \sum_{s_{h+1}\in\mathcal{S}_{h+1}^{\text{low}}} \widehat{\mathbb{T}}_h(s_{h+1} \mid s_h, a_h), \forall h \in [H], s_h \in \mathcal{S}_h^{\text{high}}, a_h \in \mathcal{A}.$$

For emission, we define

$$\widehat{\mathbb{O}}_h^{\text{trunc}}(o_h \mid s_h) := \widehat{\mathbb{O}}_h(o_h \mid s_h), \forall h \in [H], s_h \in \mathcal{S} \setminus \{s^{\text{exit}}\}, o_h \in \mathcal{O}_h,$$

while we define $\widehat{\mathbb{O}}_h^{\text{trunc}}(o_h^{\text{exit}} \mid s_h^{\text{exit}}) := 1$.

Now we define $\widehat{\mathbb{O}}_h^{\text{sub}} \in \mathbb{R}^{(|\mathcal{S}_h^{\text{high}}|+1)\times O}$ to be the sub-matrix of $\widehat{\mathbb{O}}_h^{\text{trunc}}$, where we only keep those rows of $\widehat{\mathbb{O}}_h^{\text{trunc}}$ that correspond to states from $\mathcal{S}_h^{\text{high}} \cup \{s^{\text{exit}}\}$. Similarly, we define $\mathbb{O}_h^{\text{sub}} \in \mathbb{R}^{(|\mathcal{S}_h^{\text{high}}|+1)\times O}$ to be the sub-matrix of $\mathbb{O}_h$, where we only keep those rows of $\mathbb{O}_h$ that correspond to states from $\mathcal{S}_h^{\text{high}} \cup \{s^{\text{exit}}\}$. It is easy to see that $\mathbb{O}^{\text{sub}}$ is still γ-observable. Meanwhile, we notice that

$$\begin{aligned}
\|\widehat{\mathbb{O}}_h^{\text{sub}} - \mathbb{O}_h^{\text{sub}}\|_\infty &= \max_{s_h\in\mathcal{S}_h^{\text{high}}} \|\mathbb{O}_h(\cdot \mid s_h) - \widehat{\mathbb{O}}_h(\cdot \mid s_h)\|_1 \\
&\le \max_{s_h\in\mathcal{S}_h^{\text{high}}} \sqrt{\frac{O\log(SH/\delta)}{N_h(s_h)}} \\
&\le \max_{s_h\in\mathcal{S}_h^{\text{high}}} \sqrt{\frac{2O\log(SH/\delta)}{N\epsilon_1}},
\end{aligned}$$

where the first inequality is by Lemma J.9, and the second inequality is by the definition of $\mathcal{S}_h^{\text{high}}$. Therefore, if we take

$$N \ge \frac{8O\log(SH/\delta)}{\gamma^2\epsilon_1},$$

it is guaranteed that $\|\widehat{\mathbb{O}}_h^{\text{sub}} - \mathbb{O}_h^{\text{sub}}\|_\infty \le \frac{\gamma}{2}$. Therefore, we conclude that $\widehat{\mathbb{O}}_h^{\text{sub}}$ is also $\gamma/2$-observable.

Now we are ready to examine $\widehat{\boldsymbol{b}}_h'^{,\text{trunc}}$. We firstly define the following POMDP $\widehat{\mathcal{P}}^{\text{sub}}$, which essentially deletes all states in $\mathcal{S}_h^{\text{low}} \setminus \{s^{\text{exit}}\}$ from the state space of $\widehat{\mathcal{P}}^{\text{trunc}}$ at each step h. Then we can notice that the emission of $\widehat{\mathcal{P}}^{\text{sub}}$ is exactly $\widehat{\mathbb{O}}_h^{\text{sub}}$, implying that $\widehat{\mathcal{P}}^{\text{sub}}$ is an $\gamma/2$-observable POMDP. Therefore, by [21], it is guaranteed that for any $\pi \in \Pi$,

$$\mathbb{E}_\pi^{\widehat{\mathcal{P}}^{\text{sub}}} \|\widehat{\boldsymbol{b}}_h^{\text{sub}}(\tau_h) - \widehat{\boldsymbol{b}}_h'^{,\text{sub}}(z_h)\|_1 \le \epsilon,$$

where $\widehat{\boldsymbol{b}}_h^{\text{sub}}(\tau_h), \widehat{\boldsymbol{b}}_h'^{,\text{sub}}(z_h) \in \Delta(\mathcal{S}_h^{\text{high}} \cup \{s^{\text{exit}}\})$. Now we claim that

$$\mathbb{E}_{\pi}^{\widehat{\mathcal{P}}^{\text{trunc}}} \|\widehat{\boldsymbol{b}}_h^{\text{trunc}}(\tau_h) - \boldsymbol{b}_h^{\text{apx}}(z_h)\|_1 \leq \epsilon,$$

where we define $\boldsymbol{b}_h^{\text{apx}}(z_h) \in \Delta(\mathcal{S})$ by *augmenting* $\widehat{\boldsymbol{b}}_h'^{,\text{sub}}(z_h)$ with 0 for states from $\mathcal{S}_h^{\text{low}} \setminus \{s^{\text{exit}}\}$. To see the reason, we notice that simulating $\widehat{\mathcal{P}}^{\text{trunc}}$ is exactly equivalent to simulating $\widehat{\mathcal{P}}^{\text{sub}}$, and that $\widehat{\boldsymbol{b}}_h^{\text{trunc}}(\tau_h)(s_h) = 0$ for $s_h \in \mathcal{S}_h^{\text{low}} \setminus \{s^{\text{exit}}\}$, $\widehat{\boldsymbol{b}}_h^{\text{trunc}}(\tau_h)(s_h) = \widehat{\boldsymbol{b}}_h^{\text{sub}}(\tau_h)$ for $s_h \in \mathcal{S}_h^{\text{high}} \cup \{s^{\text{exit}}\}$

For the total variation distance of trajectory distributions between $\mathcal{P}^{\text{trunc}}$ and $\widehat{\mathcal{P}}^{\text{trunc}}$, it holds that by Lemma H.7

$$\begin{aligned}
&\sum_{\overline{\tau}_H} |\mathbb{P}^{\pi,\mathcal{P}^{\text{trunc}}}(\overline{\tau}_H) - \mathbb{P}^{\pi,\widehat{\mathcal{P}}^{\text{trunc}}}(\overline{\tau}_H)| \\
&\leq \mathbb{E}_{\pi}^{\mathcal{P}^{\text{trunc}}} \sum_{h\in[H]} \|\mathbb{T}_h^{\text{trunc}}(\cdot \mid s_h, a_h) - \widehat{\mathbb{T}}_h^{\text{trunc}}(\cdot \mid s_h, a_h)\|_1 + \|\mathbb{O}_h^{\text{trunc}}(\cdot \mid s_h) - \widehat{\mathbb{O}}_h^{\text{trunc}}(\cdot \mid s_h)\|_1 \\
&= \sum_{h\in[H]} \mathbb{E}_{\pi}^{\mathcal{P}^{\text{trunc}}} \mathbb{1}[\forall t \in [h] : s_t \in \mathcal{S}_t^{\text{high}}] \|\mathbb{T}_h(\cdot \mid s_h, a_h) - \widehat{\mathbb{T}}_h(\cdot \mid s_h, a_h)\|_1 + \|\mathbb{O}_h(\cdot \mid s_h) - \widehat{\mathbb{O}}_h(\cdot \mid s_h)\|_1,
\end{aligned}$$

where the last step is due to the fact that if there exists some $t \in [h]$ such that $s_t \notin \mathcal{S}_t^{\text{high}}$, it is guaranteed that $s_h = s^{\text{exit}}$, due to the transition of $\mathcal{P}^{\text{trunc}}$, which implies that

$$\|\mathbb{T}_h^{\text{trunc}}(\cdot \mid s_h, a_h) - \widehat{\mathbb{T}}_h^{\text{trunc}}(\cdot \mid s_h, a_h)\|_1 + \|\mathbb{O}_h^{\text{trunc}}(\cdot \mid s_h) - \widehat{\mathbb{O}}_h^{\text{trunc}}(\cdot \mid s_h)\|_1 = 0.$$

We notice that for any trajectory $\overline{\tau}_h$ such that $s_t \in \mathcal{S}_t^{\text{high}}$, we have

$$\mathbb{P}^{\pi,\mathcal{P}}(\overline{\tau}_h) = \mathbb{P}^{\pi,\mathcal{P}^{\text{trunc}}}(\overline{\tau}_h),$$

which implies that

$$\begin{aligned}
&\sum_h \mathbb{E}_{\pi}^{\mathcal{P}^{\text{trunc}}} \mathbb{1}[\forall t \in [h] : s_t \in \mathcal{S}_t^{\text{high}}] \|\mathbb{T}_h(\cdot \mid s_h, a_h) - \widehat{\mathbb{T}}_h(\cdot \mid s_h, a_h)\|_1 + \|\mathbb{O}_h(\cdot \mid s_h) - \widehat{\mathbb{O}}_h(\cdot \mid s_h)\|_1 \\
&= \sum_h \mathbb{E}_{\pi}^{\mathcal{P}} \mathbb{1}[\forall t \in [h] : s_t \in \mathcal{S}_t^{\text{high}}] \|\mathbb{T}_h(\cdot \mid s_h, a_h) - \widehat{\mathbb{T}}_h(\cdot \mid s_h, a_h)\|_1 + \|\mathbb{O}_h(\cdot \mid s_h) - \widehat{\mathbb{O}}_h(\cdot \mid s_h)\|_1 \\
&\leq \sum_h \mathbb{E}_{\pi}^{\mathcal{P}} \|\mathbb{T}_h(\cdot \mid s_h, a_h) - \widehat{\mathbb{T}}_h(\cdot \mid s_h, a_h)\|_1 + \|\mathbb{O}_h(\cdot \mid s_h) - \widehat{\mathbb{O}}_h(\cdot \mid s_h)\|_1 \\
&\leq \frac{\epsilon}{3},
\end{aligned}$$

where the last step is by Theorem H.4. Hence, by Lemma H.7, it holds that

$$\mathbb{E}_{\pi}^{\mathcal{P}^{\text{trunc}}} \|\boldsymbol{b}_h^{\text{trunc}}(\tau_h) - \widehat{\boldsymbol{b}}_h^{\text{trunc}}(\tau_h)\|_1 \leq \frac{\epsilon}{3}.$$

Finally, we are ready to prove

$$\begin{aligned}
\mathbb{E}_{\pi}^{\mathcal{P}} \|\boldsymbol{b}_h(\tau_h) - \widehat{\boldsymbol{b}}_h'^{,\text{trunc}}(z_h)\|_1 &\leq \mathbb{E}_{\pi}^{\mathcal{P}} \|\boldsymbol{b}_h(\tau_h) - \boldsymbol{b}_h^{\text{trunc}}(\tau_h)\|_1 + \mathbb{E}_{\pi}^{\mathcal{P}} \|\boldsymbol{b}_h^{\text{trunc}}(\tau_h) - \widehat{\boldsymbol{b}}_h^{\text{trunc}}(\tau_h)\|_1 + \mathbb{E}_{\pi}^{\mathcal{P}} \|\widehat{\boldsymbol{b}}_h^{\text{trunc}}(\tau_h) - \boldsymbol{b}_h^{\text{apx}}(z_h)\|_1 \\
&\leq \mathbb{E}_{\pi}^{\mathcal{P}} \|\boldsymbol{b}_h(\tau_h) - \boldsymbol{b}_h^{\text{trunc}}(\tau_h)\|_1 + \mathbb{E}_{\pi}^{\mathcal{P}^{\text{trunc}}} \|\boldsymbol{b}_h^{\text{trunc}}(\tau_h) - \widehat{\boldsymbol{b}}_h^{\text{trunc}}(\tau_h)\|_1 + \mathbb{E}_{\pi}^{\widehat{\mathcal{P}}^{\text{trunc}}} \|\widehat{\boldsymbol{b}}_h^{\text{trunc}}(\tau_h) - \boldsymbol{b}_h^{\text{apx}}(z_h)\|_1 \\
&\quad + 4\|\mathbb{P}^{\pi,\mathcal{P}} - \mathbb{P}^{\pi,\mathcal{P}^{\text{trunc}}}\|_1 + 2\|\mathbb{P}^{\pi,\mathcal{P}^{\text{trunc}}} - \mathbb{P}^{\pi,\widehat{\mathcal{P}}^{\text{trunc}}}\| \\
&\leq 4HS\epsilon_1 + \epsilon_3 + \epsilon_2 + 8HS\epsilon_1 + 2\epsilon_3 \\
&\leq 12HS\epsilon_1 + \epsilon_2 + 3H\epsilon_3.
\end{aligned}$$

Therefore, for any $\epsilon > 0$, by taking $\epsilon_1 = \frac{\epsilon}{36HS}$, $\epsilon_2 = \frac{\epsilon}{3}$, $\epsilon_3 = \frac{\epsilon}{3H}$, we prove our lemma. Observe that our algorithm needed no further samples. The computational complexity follows by computing belief $\boldsymbol{b}_h^{\text{apx}}$ on POMDP $\widehat{\mathcal{P}}^{\text{sub}}$ using finite-memory policies of size $\widetilde{\Theta}(\gamma^{-4}\log(S/\epsilon))$. For the final sample complexity, we only need to ensure $N \geq \mathcal{O}(\frac{O\log(SH/\delta)}{\gamma^2\epsilon_1})$ in Equation (H.3), thus concluding our proof $\square$

H.1 Supporting Technical Lemmas

In the following, we provide some technical lemmas and their proofs.

Lemma H.6. Fix two finite sets $\mathcal{X}, \mathcal{Y}$ and two joint distributions $P_1, P_2 \in \Delta(\mathcal{X} \times \mathcal{Y})$. It holds that

$$\begin{aligned}-\mathbb{E}_{P_1(x)} \sum_y |P_1(y \,|\, x) - P_2(y \,|\, x)| &\leq \sum_{x,y} |P_1(x,y) - P_2(x,y)| - \sum_x |P_1(x) - P_2(x)| \\ &\leq \mathbb{E}_{P_1(x)} \sum_y |P_1(y \,|\, x) - P_2(y \,|\, x)|.\end{aligned}$$

Proof. For the second inequality, it holds that

$$\begin{aligned}\sum_{x,y} |P_1(x,y) - P_2(x,y)| &= \sum_{x,y} |P_1(x,y) - P_1(x)P_2(y \,|\, x) + P_1(x)P_2(y \,|\, x) - P_2(x,y)| \\ &\leq \sum_{x,y} |P_1(x,y) - P_1(x)P_2(y \,|\, x)| + |P_1(x)P_2(y \,|\, x) - P_2(x,y)| \\ &= \sum_{x,y} |P_1(x)(P_1(y \,|\, x) - P_2(y \,|\, x))| + |P_2(y \,|\, x)(P_1(x) - P_2(x))| \\ &= \mathbb{E}_{P_1(x)} \sum_y |P_1(y \,|\, x) - P_2(y \,|\, x)| + \sum_x |P_1(x) - P_2(x)|.\end{aligned}$$

Meanwhile, for the first inequality, it holds that

$$\begin{aligned}\sum_{x,y} |P_1(x,y) - P_2(x,y)| &= \sum_{x,y} |P_1(x,y) - P_1(x)P_2(y \,|\, x) + P_1(x)P_2(y \,|\, x) - P_2(x,y)| \\ &\geq \sum_{x,y} -|P_1(x,y) - P_1(x)P_2(y \,|\, x)| + |P_1(x)P_2(y \,|\, x) - P_2(x,y)| \\ &= \sum_{x,y} -|P_1(x)(P_1(y \,|\, x) - P_2(y \,|\, x))| + |P_2(y \,|\, x)(P_1(x) - P_2(x))| \\ &= \mathbb{E}_{P_1(x)} - \sum_y |P_1(y \,|\, x) - P_2(y \,|\, x)| + \sum_x |P_1(x) - P_2(x)|,\end{aligned}$$

concluding our lemma. □

Lemma H.7. Consider any two POMDP instances $\mathcal{P}$ and $\widehat{\mathcal{P}}$ and define the belief functions as $\boldsymbol{b}_{1:H}$ and $\widehat{\boldsymbol{b}}_{1:H}$, respectively (see Appendix C for definition of belief function $\boldsymbol{b}_{1:H}$). It holds that for any $\pi \in \Pi^{\text{gen}}$

$$\sum_{\overline{\tau}_H} |\mathbb{P}^{\pi,\mathcal{P}}(\overline{\tau}_H) - \mathbb{P}^{\pi,\widehat{\mathcal{P}}}(\overline{\tau}_H)| \leq \mathbb{E}^{\mathcal{P}}_{\pi} \sum_{h \in [H-1]} \|\mathbb{T}_h(\cdot \,|\, s_h, a_h) - \widehat{\mathbb{T}}_h(\cdot \,|\, s_h, a_h)\|_1 + \mathbb{E}^{\mathcal{P}}_{\pi} \sum_{h \in [H]} \|\mathbb{O}_h(\cdot \,|\, s_h) - \widehat{\mathbb{O}}_h(\cdot \,|\, s_h)\|_1.$$

$$\mathbb{E}^{\mathcal{P}}_{\pi} \|\boldsymbol{b}_H(\tau_H) - \widehat{\boldsymbol{b}}_H(\tau_H)\|_1 \leq 2\mathbb{E}^{\mathcal{P}}_{\pi} \sum_{h \in [H-1]} \|\mathbb{T}_h(\cdot \,|\, s_h, a_h) - \widehat{\mathbb{T}}_h(\cdot \,|\, s_h, a_h)\|_1 + 2\mathbb{E}^{\mathcal{P}}_{\pi} \sum_{h \in [H]} \|\mathbb{O}_h(\cdot \,|\, s_h) - \widehat{\mathbb{O}}_h(\cdot \,|\, s_h)\|_1,$$

where we remind readers that we denote by $\overline{\tau}_H = (s_{1:H}, o_{1:H}, a_{1:H-1})$ the trajectory with states from an episode of the POMDP, and w.l.o.g, we assume $\widehat{\mathbb{T}}_H = \mathbb{T}_H$ for notational convenience, since one can assume the episode ends deterministically at a state s_{H+1}.

Proof. The first inequality also appears in [36] and we provide a simplified proof here for completeness. By Lemma H.6, it holds that

$$\begin{aligned}
&\sum_{\overline{\tau}_H} |\mathbb{P}^{\pi,\mathcal{P}}(\overline{\tau}_H) - \mathbb{P}^{\pi,\widehat{\mathcal{P}}}(\overline{\tau}_H)| \le \sum_{\overline{\tau}_{H-1}} |\mathbb{P}^{\pi,\mathcal{P}}(\overline{\tau}_{H-1}) - \mathbb{P}^{\pi,\widehat{\mathcal{P}}}(\overline{\tau}_{H-1})| \\
&+ \mathbb{E}_\pi^{\mathcal{P}} \sum_{a_{H-1},s_H,o_H} \Big| \pi_{H-1}(a_{h-1} \,|\, \overline{\tau}_{H-1}) \mathbb{T}_{H-1}(s_H \,|\, s_{H-1}, a_{H-1}) \mathbb{O}_H(o_H \,|\, s_H) \\
&\qquad\qquad - \pi_{H-1}(a_{h-1} \,|\, \overline{\tau}_{H-1}) \widehat{\mathbb{T}}_{H-1}(s_H \,|\, s_{H-1}, a_{H-1}) \widehat{\mathbb{O}}_H(o_H \,|\, s_H) \Big| \\
&\le \sum_{\overline{\tau}_{H-1}} |\mathbb{P}^{\pi,\mathcal{P}}(\overline{\tau}_{H-1}) - \mathbb{P}^{\pi,\widehat{\mathcal{P}}}(\overline{\tau}_{H-1})| + \mathbb{E}_\pi^{\mathcal{P}} \|\mathbb{T}_{H-1}(\cdot \,|\, s_{H-1}, a_{H-1}) - \widehat{\mathbb{T}}_{H-1}(\cdot \,|\, s_{H-1}, a_{H-1})\|_1 \\
&\quad + \|\mathbb{O}_H(\cdot \,|\, s_H) - \widehat{\mathbb{O}}_H(\cdot \,|\, s_H)\|_1,
\end{aligned}$$

where the last step is again from Lemma H.6. Therefore, by rolling out the inequality repeatedly, we proved the first result.

For the second result, we notice that by Lemma H.6, it holds that

$$\sum_{\tau_h,s_h} |\mathbb{P}_h^{\pi,\mathcal{P}}(\tau_h, s_h) - \mathbb{P}_h^{\pi,\widehat{\mathcal{P}}}(\tau_h, s_h)| \ge -\sum_{\tau_h} |\mathbb{P}_h^{\pi,\mathcal{P}}(\tau_h) - \mathbb{P}_h^{\pi,\widehat{\mathcal{P}}}(\tau_h)| + \mathbb{E}_\pi^{\mathcal{P}} \sum_{s_h} |\mathbb{P}_h^{\pi,\mathcal{P}}(s_h \,|\, \tau_h) - \mathbb{P}_h^{\pi,\widehat{\mathcal{P}}}(s_h \,|\, \tau_h)|.$$

Notice the fact that $\mathbb{P}_h^{\pi,\mathcal{P}}(s_h \,|\, \tau_h)$ does not depend on π since it is exactly $\boldsymbol{b}_h(\tau_h)(s_h)$, we conclude that

$$\begin{aligned}
\mathbb{E}_\pi^{\mathcal{P}} \|\boldsymbol{b}_h(\tau_h) - \widehat{\boldsymbol{b}}_h(\tau_h)\|_1 \le& \sum_{\tau_h,s_h} |\mathbb{P}_h^{\pi,\mathcal{P}}(\tau_h, s_h) - \mathbb{P}_h^{\pi,\widehat{\mathcal{P}}}(\tau_h, s_h)| + \sum_{\tau_h} |\mathbb{P}_h^{\pi,\mathcal{P}}(\tau_h) - \mathbb{P}_h^{\pi,\widehat{\mathcal{P}}}(\tau_h)| \\
\le& 2 \sum_{\overline{\tau}_H} |\mathbb{P}^{\pi,\mathcal{P}}(\overline{\tau}_h) - \mathbb{P}^{\pi,\widehat{\mathcal{P}}}(\overline{\tau}_h)|,
\end{aligned}$$

where the last step comes from the fact that after marginalization, the total variation distance will not increase. By combining it with the first result, we proved the second result. □

Lemma H.8. For any reward function $r_{1:H}$ of $\mathcal{P}$ with $r_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ for any $h \in [H]$, it holds that

$$\max_{\pi \in \Pi^{\text{gen}}} v^{\mathcal{P}}(\pi) \le \max_{\pi \in \Pi_{\mathcal{S}}} v^{\mathcal{P}}(\pi)$$

Proof. Denote $\pi^\star \in \Pi_{\mathcal{S}}$ to be the optimal policy obtained by running value iteration only on the state space for $\mathcal{P}$. Now we are ready to prove the following argument for any $\pi \in \Pi^{\text{gen}}$ inductively:

$$Q_h^{\pi^\star,\mathcal{P}}(s_h, a_h) \ge Q_h^{\pi,\mathcal{P}}(s_{1:h}, o_{1:h}, a_{1:h}).$$

It is easy to see the argument holds for $h = H$. Fix state-action pair $(s_h, a_h) \in \mathcal{S} \times \mathcal{A}$ and trajectory $(s_{1:h-1}, o_{1:h}, a_{1:h-1})$, we note that:

$$\begin{aligned}
Q_h^{\pi^\star,\mathcal{P}}(s_h, a_h) &= r_h(s_h, a_h) + \mathbb{E}_{s_{h+1} \sim \mathbb{T}_h(\cdot \,|\, s_h, a_h)} \left[\max_{a_{h+1}} Q_{h+1}^{\pi^\star,\mathcal{P}}(s_{h+1}, a_{h+1}) \right] \\
&\ge r_h(s_h, a_h) + \mathbb{E}_{\substack{s_{h+1} \sim \mathbb{T}_h(\cdot \,|\, s_h, a_h), \\ o_{h+1} \sim \mathbb{O}_{h+1}(\cdot | s_{h+1})}} \left[\max_{a_{h+1}} Q_{h+1}^{\pi,\mathcal{P}}(s_{1:h+1}, o_{1:h+1}, a_{1:h+1}) \right] \\
&\ge r_h(s_h, a_h) + \mathbb{E}_{\substack{s_{h+1} \sim \mathbb{T}_h(\cdot \,|\, s_h, a_h), \\ o_{h+1} \sim \mathbb{O}_{h+1}(\cdot | s_{h+1})}} \left[\mathbb{E}_{a_{h+1} \sim \pi_h(\cdot \,|\, s_{1:h+1}, o_{1:h+1}, a_{1:h})} Q_{h+1}^{\pi,\mathcal{P}}(s_{1:h+1}, o_{1:h+1}, a_{1:h+1}) \right] \\
&= Q_h^{\pi,\mathcal{P}}(s_{1:h}, o_{1:h}, a_{1:h}),
\end{aligned}$$

where the first inequality comes from the inductive hypothesis. □

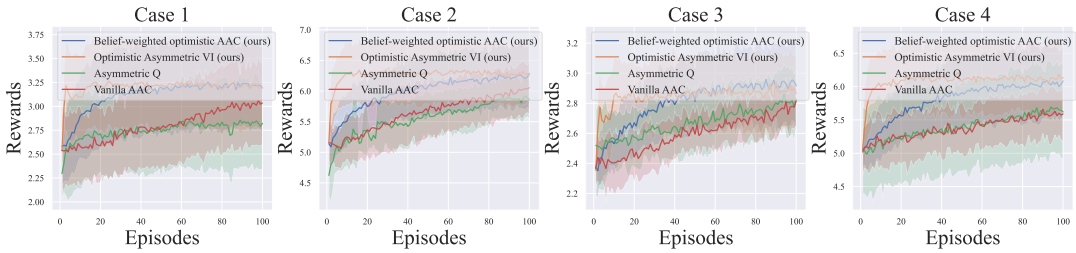

Figure 2: Results for POMDPs of different sizes, where our methods achieve the best performance with the smallest samples (VI: value iteration; AAC: asymmetric actor-critic).

		Asymmetric optimistic NPG	Expert policy distillation	Asymmetric Q learning	Vanilla AAC
Deterministic POMDP	Case 1	3.32 ± 0.66	$\mathbf{3.33 \pm 0.62}$	3.04 ± 0.58	3.25 ± 0.65
	Case 2	7.1 ± 0.48	$\mathbf{7.26 \pm 0.68}$	6.15 ± 0.85	6.41 ± 0.91
	Case 3	3.04 ± 0.23	$\mathbf{3.25 \pm 0.33}$	3.09 ± 0.39	3.1 ± 0.38
	Case 4	6.51 ± 0.6	$\mathbf{6.54 \pm 0.58}$	6.16 ± 0.48	5.87 ± 0.58
Block MDP	Case 1	3.31 ± 0.46	$\mathbf{3.37 \pm 0.41}$	3.03 ± 0.4	3.08 ± 0.43
	Case 2	6.36 ± 0.52	$\mathbf{6.67 \pm 0.54}$	5.74 ± 0.43	5.7 ± 0.46
	Case 3	3.2 ± 0.26	$\mathbf{3.37 \pm 0.22}$	3.14 ± 0.31	2.97 ± 0.32
	Case 4	6.01 ± 0.32	$\mathbf{6.44 \pm 0.36}$	5.58 ± 0.32	5.33 ± 0.19

Table 2: Rewards of different approaches for POMDPs under the deterministic filter condition.

I Missing Details in Section 6

POMDP under deterministic filter condition. We first evaluate our algorithms on POMDPs with certain structures, i.e., the deterministic conditions. In particular, we generate POMDPs, where either the transition dynamics are deterministic or the emission ensures decodability. We test three of our approaches, expert policy distillation, asymmetric optimistic natural policy gradient. We summarize our results in Table 2, where the four cases corresponds to POMDPs with $(S = A = 2, O = 3, H = 5)$, $(S = A = 2, O = 3, H = 10)$, $(S = 3, A = 2, O = 4, H = 5)$, $(S = 3, A = 2, O = 4, H = 5)$, and we can see that our approach based on expert distillation outperforms all the other methods, which is consistent with the fact that such methods have exploited the special structures of the POMDPs achieving both polynomial sample and computation complexity.

General POMDPs. Here we also evaluate our methods for general randomly generated POMDPs without any structures. Hence, we compare the baselines with asymmetric optimistic natural policy gradient and asymmetric optimistic value iteration (i.e., the single-agent version of Algorithm 5). In Figure 2, we show the performance of different algorithms in POMDPs of different size, where the four cases corresponds to POMDPs with $(S = A = O = 2, H = 5)$, $(S = A = O = 2, H = 10)$, $(S = O = 3, A = 2, H = 10)$, $(S = A = 3, O = 2, H = 10)$, and our approaches achieves the highest rewards with small number of episodes.

Implementation details. For each problem setting, we generated 20 POMDPs randomly and report the average performance and its standard deviation for each algorithms. For our algorithms based on privileged value learning methods, we find that using a finite memory of 3 already provides strong performance. For our algorithms based privileged policy learning, we instantiate the MDP learning algorithm with the fully observable optimistic natural policy gradient algorithm [65]. Meanwhile, for both the decoder learning and belief learning, we find that just utilizing all the historic trajectories gives us reasonable performance without additional samples. For baselines, the hyperparameters α for Q-value update and step size for the policy update are tuned by grid search, where α controls the update of temporal difference learning (recall the update rule of temporal difference learning as $Q \leftarrow (1-\alpha)Q + \alpha Q^{\text{target}}$). For asymmetric Q-learning, we use ϵ-greedy exploration, where we use the seminal decreasing rate $\epsilon_t = \frac{H+1}{H+t}$. Finally, all simulations are conducted on a personal laptop with Apple M1 CPU and 16 GB memory.

Empirical insights and interpretation of the experimental results. To understand intuitively why our approach outperforms those baseline algorithms, we notice the key difference in the value and policy update style between our approaches and vanilla asymmetric actor critic and asymmetric Q-learning. The baselines often roll-out the polices, collect trajectories, and only update the value and the policies on the states/history the trajectories have visited. Therefore, ideally, to learn a good policy for baselines, the number of trajectories to collect is at least as large as the history size, which is indeed exponential in the horizon H. This is known as curse of history for partially observable RL. In contrast, in our algorithms, we explicitly estimate the empirical transition and emissions, which is indeed of polynomial size. Thus, the sample complexity avoids suffering from the potential exponential dependency of horizon or the length of the finite memory. Finally, we acknowledge that the baselines are developed to handle complex, high-dimensional deep RL problems, while scaling our methods to deep RL benchmarks requires non-trivial engineering efforts.

J Missing Details in Section 7

Proof of Proposition 7.1:

Given the conditions of Definition 3.2, the function ϕ_h that satisfies the condition of Proposition 7.1 can be constructed recursively as follows

$$\phi_h(\tau_h) := \psi_h(\phi_{h-1}(\tau_{h-1}), a_{h-1}, o_h), \forall \tau_h \in \mathcal{T}_{h-1},$$

and $\phi_1(o_1) := \psi_1(o_1)$. Therefore, we can prove by induction that by belief update rule

$$\mathbb{P}^{\mathcal{P}}(s_h \,|\, \tau_h) = \boldsymbol{b}_h(b^{\phi_{h-1}(\tau_{h-1})}, a_{h-1}, o_h) = b^{\psi_h(\phi_{h-1}(\tau_{h-1}), a_{h-1}, o_h)},$$

where the last step is by the definition of our ϕ_h. Therefore, we have $\mathbb{P}^{\mathcal{P}}(s_h = \psi_h(\phi_{h-1}(\tau_{h-1}), a_{h-1}, o_h) \,|\, \tau_h) = 1$.

For the other direction, it is similar to the proof of Lemma C.1 in [14] for m-step decodable POMDP. Here we prove it for completeness. For any trajectory $\tau_h \in \mathcal{T}_h$, it holds by the belief update that

$$\mathbb{P}^{\mathcal{P}}(s_h \,|\, \tau_h) = \boldsymbol{b}_h(b^{\phi_{h-1}(\tau_{h-1})}, a_{h-1}, o_h) = \mathbb{P}^{\mathcal{P}}(s_h \,|\, s_{h-1} = \phi_{h-1}(\tau_{h-1}), a_{h-1}, o_h).$$

Meanwhile, since we know $\mathbb{P}^{\mathcal{P}}(\cdot \,|\, \tau_h)$ is a one-hot vector, we can construct ψ_h such that $\psi_h(s_{h-1}, a_{h-1}, o_h)$ is the unique s_h that makes $\mathbb{P}^{\mathcal{P}}(s_h \,|\, \tau_h) > 0$ with $s_{h-1} = \phi_{h-1}(\tau_{h-1})$. If this procedure does not complete the definition of ψ for some s_{h-1}, a_{h-1}, o_h, it implies that either s_{h-1} is not reachable or o_h is not reachable given s_{h-1}, i.e., $\mathbb{P}^{\mathcal{P}}(o_h \,|\, s_{h-1}, a'_{h-1}) = 0$ for any $a'_{h-1} \in \mathcal{A}$, thus recovering the conditions of Definition 3.2. ■

Proof of Proposition 7.3: Note that for any given problem instance of a POMDP, we can add a dummy agent that has the observation being the exact state at each timestep, and has only one dummy action that does not affect the transition or reward. Therefore, even the local private information $p_{i,h}$ of the dummy agent can decode the latent state, and hence c_h, p_h reveals the latent state.Therefore, the corresponding POSG with the dummy agent satisfies the condition in Proposition 7.3. Meanwhile, it is direct to see that any CCE of the POSG is an optimal policy for the original POMDP. Now by the PSPACE-hardness of planning for POMDPs [57], we proved our proposition. ■

Theorem J.1. For any $\pi \in \Pi_{\mathcal{S}}$ and (potentially stochastic) decoding functions $\widehat{g} = \{\widehat{g}_{i,h}\}_{i\in[n],h\in[H]}$ with $\widehat{g}_{i,h} : \mathcal{C}_h \times \mathcal{P}_{i,h} \to \Delta(\mathcal{S})$ for each $i \in [n], h \in [H]$, it holds that

$$\text{NE/CCE-gap}(\pi^{\widehat{g}}) - \text{NE/CCE-gap}(\pi) \le 2nH^2 \max_{i\in[n]} \max_{u_i\in\Pi_i} \max_{j\in[n],h\in[H]} \mathbb{P}^{u_i\times\pi_{-i},\mathcal{G}}(s_h \neq \widehat{g}_{j,h}(c_h, p_{j,h}))$$

$$\text{CE-gap}(\pi^{\widehat{g}}) - \text{CE-gap}(\pi) \le 2nH^2 \max_{i\in[n]} \max_{m_i\in\mathcal{M}_i} \max_{j\in[n],h\in[H]} \mathbb{P}^{(m_i\diamond\pi_i)\odot\pi_{-i},\mathcal{G}}(s_h \neq \widehat{g}_{j,h}(c_h, p_{j,h}))$$

where $\pi^{\widehat{g}}$ is the distilled policy of π through the decoding functions $\widehat{g}$, where at step h, each agent i firstly individually decodes the state by sampling $s_h \sim \widehat{g}_{i,h}(\cdot \,|\, c_h, p_{i,h})$, and then act according to the expert $\pi_{i,h}$. In other words, the decoding process does not need correlations among the agents.

Proof. For notational simplicity, we write v_i instead of $v_i^{\mathcal{G}}$ when the underlying model is clear in the context. Firstly, we consider any deterministic decoding function $\widehat{\phi} = \{\widehat{\phi}_{i,h}\}_{i\in[n],h\in[H]}$ with

$\widehat{\phi}_{i,h} : \mathcal{C}_h \times \mathcal{P}_{i,h} \to \mathcal{S}$ for each $i \in [n], h \in [H]$, and note the following that for any $i \in [n]$,

$$\begin{aligned}
v_i(\pi) - v_i(\pi^{\widehat{\phi}}) &= \mathbb{E}^{\mathcal{G}}_{\pi}[R] - \mathbb{E}^{\mathcal{G}}_{\pi^{\widehat{\phi}}}[R] \\
&= \mathbb{E}^{\mathcal{G}}_{\pi}[R\mathbb{1}[\forall j \in [n], h \in [H] : s_h = \widehat{\phi}_{j,h}(c_h, p_{j,h})]] - \mathbb{E}^{\mathcal{G}}_{\pi^{\widehat{\phi}}}[R\mathbb{1}[\forall j \in [n], h \in [H] : s_h = \widehat{\phi}_{j,h}(c_h, p_{j,h})]] \\
&\quad + \mathbb{E}^{\mathcal{G}}_{\pi}[R\mathbb{1}[\exists j \in [n], h \in [H] : s_h \neq \widehat{\phi}_{j,h}(c_h, p_{j,h})]] - \mathbb{E}^{\mathcal{G}}_{\pi^{\widehat{\phi}}}[R\mathbb{1}[\exists j \in [n], h \in [H] : s_h \neq \widehat{\phi}_{j,h}(c_h, p_{j,h})]] \\
&= \mathbb{E}^{\mathcal{G}}_{\pi}[R\mathbb{1}[\exists j \in [n], h \in [H] : s_h \neq \widehat{\phi}_{j,h}(c_h, p_{j,h})]] - \mathbb{E}^{\mathcal{G}}_{\pi^{\widehat{\phi}}}[R\mathbb{1}[\exists j \in [n], h \in [H] : s_h \neq \widehat{\phi}_{j,h}(c_h, p_{j,h})]],
\end{aligned}$$

where we define $R := \sum_h r_{i,h}(s_h, a_h)$ and the last step is by the definition of $\pi^{\widehat{\phi}}$. Therefore, we conclude that

$$v_i(\pi) - v_i(\pi^{\widehat{\phi}}) \leq H\mathbb{P}^{\pi,\mathcal{G}}(\exists j \in [n], h \in [H] : s_h \neq \widehat{\phi}_{j,h}(c_h, p_{j,h})).$$

Therefore, by noting the fact that $\widehat{g}$ is equivalent to a mixture of deterministic decoding functions, where at the beginning of each episode, one can firstly independently sample the outcome for each $c_h, p_{j,h}$ for $j \in [n]$ and $h \in [H]$, we conclude that

$$\begin{aligned}
v_i(\pi) - v_i(\pi^{\widehat{g}}) &= v_i(\pi) - \mathbb{E}_{\widehat{\phi}\sim\widehat{g}} v_i(\pi^{\widehat{\phi}}) \\
&\leq H\mathbb{E}_{\widehat{\phi}\sim\widehat{g}}\mathbb{P}^{\pi,\mathcal{G}}(\exists j \in [n], h \in [H] : s_h \neq \widehat{\phi}_{j,h}(c_h, p_{j,h})) \\
&= H\mathbb{P}^{\pi,\mathcal{G}}(\exists j \in [n], h \in [H] : s_h \neq \widehat{g}_{j,h}(c_h, p_{j,h})).
\end{aligned}$$

Now we prove our result for NE and CCE first. Due to similar arguments for evaluating $v_i(\pi) - v_i(\pi^{\widehat{\phi}})$, for any $u_i \in \Pi_i \cup \Pi_{\mathcal{S},i}$, it holds that

$$\begin{aligned}
v_i(u_i \times \pi^{\widehat{\phi}}_{-i}) - v_i(u_i \times \pi_{-i}) &\leq \mathbb{E}^{\mathcal{G}}_{u_i \times \pi^{\widehat{\phi}}_{-i}}[R\mathbb{1}[\exists j \in [n] \setminus \{i\}, h \in [H] : s_h \neq \widehat{\phi}_{j,h}(c_h, p_{j,h})]] \\
&\leq H\mathbb{P}^{u_i \times \pi^{\widehat{\phi}}_{-i},\mathcal{G}}(\exists j \in [n] \setminus \{i\}, h \in [H] : s_h \neq \widehat{\phi}_{j,h}(c_h, p_{j,h})).
\end{aligned}$$

We notice the following fact that

$$\mathbb{P}^{u_i \times \pi_{-i},\mathcal{G}}(\forall j \in [n] \setminus \{i\}, h \in [H] : s_h = \widehat{\phi}_{j,h}(c_h, p_{j,h})) = \sum_{\bar{\tau}_H \in \overline{\mathcal{T}}_H(\widehat{\phi})} \mathbb{P}^{u_i \times \pi_{-i},\mathcal{G}}(\bar{\tau}_H),$$

where we define $\overline{\mathcal{T}}_H(\widehat{\phi}) := \{\bar{\tau}_H \in \overline{\mathcal{T}}_H \,|\, \forall j \in [n] \setminus \{i\}, h \in [H] : s_h = \widehat{\phi}_{j,h}(c_h, p_{j,h})\}$. By definition of $\pi^{\widehat{\phi}}$, it holds that

$$\forall \bar{\tau}_H \in \overline{\mathcal{T}}_H(\widehat{\phi}) : \mathbb{P}^{u_i \times \pi_{-i},\mathcal{G}}(\bar{\tau}_H) = \mathbb{P}^{u_i \times \pi^{\widehat{\phi}}_{-i},\mathcal{G}}(\bar{\tau}_H).$$

Therefore, we have

$$\mathbb{P}^{u_i \times \pi_{-i},\mathcal{G}}(\forall j \in [n] \setminus \{i\}, h \in [H] : s_h = \widehat{\phi}_{j,h}(c_h, p_{j,h})) = \mathbb{P}^{u_i \times \pi^{\widehat{\phi}}_{-i},\mathcal{G}}(\forall j \in [n] \setminus \{i\}, h \in [H] : s_h = \widehat{\phi}_{j,h}(c_h, p_{j,h})).$$

Correspondingly, it holds that

$$\mathbb{P}^{u_i \times \pi_{-i},\mathcal{G}}(\exists j \in [n] \setminus \{i\}, h \in [H] : s_h \neq \widehat{\phi}_{j,h}(c_h, p_{j,h})) = \mathbb{P}^{u_i \times \pi^{\widehat{\phi}}_{-i},\mathcal{G}}(\exists j \in [n] \setminus \{i\}, h \in [H] : s_h \neq \widehat{\phi}_{j,h}(c_h, p_{j,h})),$$

which implies that

$$\begin{aligned}
v_i(u_i \times \pi^{\widehat{\phi}}_{-i}) - v_i(u_i \times \pi_{-i}) &\leq H\mathbb{P}^{u_i \times \pi^{\widehat{\phi}}_{-i},\mathcal{G}}(\exists j \in [n] \setminus \{i\}, h \in [H] : s_h \neq \widehat{\phi}_{j,h}(c_h, p_{j,h})) \\
&= H\mathbb{P}^{u_i \times \pi_{-i},\mathcal{G}}(\exists j \in [n] \setminus \{i\}, h \in [H] : s_h \neq \widehat{\phi}_{j,h}(c_h, p_{j,h})).
\end{aligned}$$

Again by the fact that $\widehat{g}$ is equivalent to a mixture of deterministic decoding functions, it holds

$$\begin{aligned}
v_i(u_i \times \pi^{\widehat{g}}_{-i}) - v_i(u_i \times \pi_{-i}) &= \mathbb{E}_{\widehat{\phi}\sim\widehat{g}} v_i(u_i \times \pi^{\widehat{\phi}}_{-i}) - v_i(u_i \times \pi_{-i}) \\
&\leq H\mathbb{E}_{\widehat{\phi}\sim\widehat{g}}\mathbb{P}^{u_i \times \pi_{-i},\mathcal{G}}(\exists j \in [n] \setminus \{i\}, h \in [H] : s_h \neq \widehat{\phi}_{j,h}(c_h, p_{j,h})) \\
&= H\mathbb{P}^{u_i \times \pi_{-i},\mathcal{G}}(\exists j \in [n] \setminus \{i\}, h \in [H] : s_h \neq \widehat{g}_{j,h}(c_h, p_{j,h})).
\end{aligned} \tag{J.1}$$

Now we are ready to evaluate the NE/CCE gap of policy $\pi^{\widehat{g}}$ as follows:

$$
\begin{aligned}
&\text{NE/CCE-gap}(\pi^{\widehat{g}}) - \text{NE/CCE-gap}(\pi) \\
&\le \max_{i\in[n]} \left(\max_{u_i\in\Pi_i} v_i(u_i \times \pi^{\widehat{g}}_{-i}) - \max_{u_i\in\Pi_{\mathcal{S},i}} v_i(u_i\times\pi_{-i}) \right) + H\mathbb{P}^{\pi,\mathcal{G}}(\exists j\in[n], h\in[H] : s_h \neq \widehat{g}_{j,h}(c_h, p_{j,h})).
\end{aligned}
$$

Now we notice that $\max_{u_i\in\Pi_{\mathcal{S},i}} v_i(u_i\times\pi_{-i}) = \max_{u_i\in\Pi_i} v_i(u_i\times\pi_{-i})$ since $\Pi_{\mathcal{S},i}\subseteq\Pi_i$ by the deterministic filter condition Definition 3.2 and π_{-i} is a Markov policy. Therefore, we conclude that

$$
\begin{aligned}
&\text{NE/CCE-gap}(\pi^{\widehat{g}}) - \text{NE/CCE-gap}(\pi) \\
&\le \max_{i\in[n]} \left(\max_{u_i\in\Pi_i} v_i(u_i \times \pi^{\widehat{g}}_{-i}) - \max_{u_i\in\Pi_i} v_i(u_i\times\pi_{-i}) \right) + H\mathbb{P}^{\pi,\mathcal{G}}(\exists j\in[n], h\in[H] : s_h \neq \widehat{g}_{j,h}(c_h, p_{j,h})) \\
&\le \max_{i\in[n]} \left(\max_{u_i\in\Pi_i} \left(v_i(u_i \times \pi^{\widehat{g}}_{-i}) - v_i(u_i\times\pi_{-i}) \right) \right) + H\mathbb{P}^{\pi,\mathcal{G}}(\exists j\in[n], h\in[H] : s_h \neq \widehat{g}_{j,h}(c_h, p_{j,h})) \\
&\le \max_{i\in[n]} \left(\max_{u_i\in\Pi_i} H\mathbb{P}^{u_i\times\pi_{-i},\mathcal{G}}(\exists j\in[n]\setminus\{i\}, h\in[H] : s_h \neq \widehat{\phi}_{j,h}(c_h, p_{j,h})) \right) \\
&\qquad + H\mathbb{P}^{\pi,\mathcal{G}}(\exists j\in[n], h\in[H] : s_h \neq \widehat{g}_{j,h}(c_h, p_{j,h})) \\
&\le 2H \max_{i\in[n], u_i\in\Pi_i} \sum_{j\in[n]}\sum_h \mathbb{P}^{u_i\times\pi_{-i},\mathcal{G}}(s_h \neq \widehat{g}_{j,h}(c_h, p_{j,h})) \\
&\le 2nH^2 \max_{i\in[n], u_i\in\Pi_i, j\in[n], h\in[H]} \mathbb{P}^{u_i\times\pi_{-i},\mathcal{G}}(s_h \neq \widehat{g}_{j,h}(c_h, p_{j,h})),
\end{aligned}
$$

where the second last step is by a union bound, thus proving our result for NE and CCE.

For CE, consider any strategy modification $m_i \in \mathcal{M}_i \cup \mathcal{M}_{\mathcal{S},i}$, it holds that

$$
\begin{aligned}
&\text{CE-gap}(\pi^{\widehat{g}}) - \text{CE-gap}(\pi) \\
&\le \max_{m_i\in\mathcal{M}_i} v_i((m_i\diamond\pi^{\widehat{g}}_i)\odot\pi^{\widehat{g}}_{-i}) - \max_{m_i\in\mathcal{M}_{\mathcal{S},i}} v_i((m_i\diamond\pi_i)\odot\pi_{-i}) + H\mathbb{P}^{\pi,\mathcal{G}}(\exists j\in[n], h\in[H] : s_h \neq \widehat{g}_{j,h}(c_h, p_{j,h})).
\end{aligned}
$$

Now we notice that $\max_{m_i\in\Pi_{\mathcal{S},i}} v_i((m_i\diamond\pi_i)\odot\pi_{-i}) = \max_{m_i\in\mathcal{M}_{\mathcal{S},i}} v_i((m_i\diamond\pi_i)\odot\pi_{-i})$ since $\mathcal{M}_{\mathcal{S},i}\subseteq\mathcal{M}_i$ by the deterministic filter condition Definition 3.2 and Lemma J.2. Therefore, we conclude that

$$
\begin{aligned}
&\text{CE-gap}(\pi^{\widehat{g}}) - \text{CE-gap}(\pi) \\
&\le \max_{i\in[n]}\max_{m_i\in\mathcal{M}_i} v_i((m_i\diamond\pi^{\widehat{g}}_i)\odot\pi^{\widehat{g}}_{-i}) - \max_{m_i\in\mathcal{M}_i} v_i((m_i\diamond\pi_i)\odot\pi_{-i}) + H\mathbb{P}^{\pi,\mathcal{G}}(\exists j\in[n], h\in[H] : s_h \neq \widehat{g}_{j,h}(c_h, p_{j,h})) \\
&\le \max_{i\in[n]}\max_{m_i\in\mathcal{M}_i} \left(v_i((m_i\diamond\pi^{\widehat{g}}_i)\odot\pi^{\widehat{g}}_{-i}) - v_i((m_i\diamond\pi_i)\odot\pi_{-i}) \right) + H\mathbb{P}^{\pi,\mathcal{G}}(\exists j\in[n], h\in[H] : s_h \neq \widehat{g}_{j,h}(c_h, p_{j,h})) \\
&\le \max_{i\in[n]}\max_{m_i\in\mathcal{M}_i} H\mathbb{P}^{(m_i\diamond\pi_i)\odot\pi_{-i},\mathcal{G}}(\exists j\in[n]\setminus\{i\}, h\in[H] : s_h \neq \widehat{\phi}_{j,h}(c_h, p_{j,h})) \\
&\qquad + H\mathbb{P}^{\pi,\mathcal{G}}(\exists j\in[n], h\in[H] : s_h \neq \widehat{\phi}_{j,h}(c_h, p_{j,h})) \\
&\le 2H \max_{i\in[n], m_i\in\mathcal{M}_i} \sum_{j\in[n]}\sum_h \mathbb{P}^{(m_i\diamond\pi_i)\odot\pi_{-i},\mathcal{G}}(s_h \neq \widehat{\phi}_{j,h}(c_h, p_{j,h})) \\
&\le 2nH^2 \max_{i\in[n], m_i\in\mathcal{M}_i, h\in[H], j\in[n]} \mathbb{P}^{(m_i\diamond\pi_i)\odot\pi_{-i},\mathcal{G}}(s_h \neq \widehat{\phi}_{j,h}(c_h, p_{j,h}))
\end{aligned}
$$

where the third step is due to the same reason as Equation (J.1). □

Lemma J.2. For any $\pi\in\Pi_{\mathcal{S}}$, it holds that for any $i\in[n]$ that

$$
\max_{m_i\in\mathcal{M}_i^{\text{gen}}} v_i((m_i\diamond\pi_i)\odot\pi_{-i}) = \max_{m_i\in\mathcal{M}_{\mathcal{S},i}} v_i((m_i\diamond\pi_i)\odot\pi_{-i}).
$$

Proof. Denote $m_i^\star \in \operatorname{argmax}_{m_i\in\mathcal{M}_i^{\text{gen}}} v_i((m_i\diamond\pi_i)\odot\pi_{-i})$ and $\widehat{m}_i^\star \in \operatorname{argmax}_{m_i\in\mathcal{M}_{\mathcal{S},i}} v_i((m_i\diamond\pi_i)\odot\pi_{-i})$. Now we shall prove that $V_{i,h}^{(m_i^\star\diamond\pi_i)\odot\pi_{-i},\mathcal{G}}(s_{1:h}, o_{1:h}, a_{1:h-1}) \le V_{i,h}^{(\widehat{m}_i^\star\diamond\pi_i)\odot\pi_{-i},\mathcal{G}}(s_h)$

inductively for each $h \in [H]$. Note that it holds for $h = H+1$. Now we consider the following

$$
\begin{aligned}
&V_{i,h}^{(m_i^\star \diamond \pi_i)\odot \pi_{-i}, \mathcal{G}}(s_{1:h}, o_{1:h}, a_{1:h-1}) \\
&= \mathbb{E}_{\substack{a_h \sim \pi_h(\cdot \mid s_h) \\ s_{h+1} \sim \mathbb{T}_h(\cdot \mid s_h, m_{i,h}^\star(s_{1:h}, o_{1:h}, a_{1:h-1}, a_{i,h}), a_{-i,h}) \\ o_{h+1} \sim \mathbb{O}_{h+1}(\cdot \mid s_{h+1})}} \Big\{ r_h(s_h, m_{i,h}^\star(s_{1:h}, o_{1:h}, a_{1:h-1}, a_{i,h}), a_{-i,h}) \\
&\qquad + V_{i,h+1}^{(m_i^\star \diamond \pi_i)\odot \pi_{-i}, \mathcal{G}}(s_{1:h+1}, o_{1:h+1}, a_{1:h-1}, m_{i,h}^\star(s_{1:h}, o_{1:h}, a_{1:h-1}, a_{i,h}), a_{-i,h}) \Big\} \\
&\le \mathbb{E}_{\substack{a_h \sim \pi_h(\cdot \mid s_h) \\ s_{h+1} \sim \mathbb{T}_h(\cdot \mid s_h, m_{i,h}^\star(s_{1:h}, o_{1:h}, a_{1:h-1}, a_{i,h}), a_{-i,h}) \\ o_{h+1} \sim \mathbb{O}_{h+1}(\cdot \mid s_{h+1})}} \Big[r_h(s_h, m_{i,h}^\star(s_{1:h}, o_{1:h}, a_{1:h-1}, a_{i,h}), a_{-i,h}) + V_{i,h+1}^{(\widehat{m}_i^\star \diamond \pi_i)\odot \pi_{-i}, \mathcal{G}}(s_{h+1}) \Big] \\
&\le \mathbb{E}_{\substack{a_h \sim \pi_h(\cdot \mid s_h) \\ s_{h+1} \sim \mathbb{T}_h(\cdot \mid s_h, \widehat{m}_{i,h}^\star(s_h, a_{i,h}), a_{-i,h}) \\ o_{h+1} \sim \mathbb{O}_{h+1}(\cdot \mid s_{h+1})}} \Big[r_h(s_h, \widehat{m}_{i,h}^\star(s_h, a_{i,h}), a_{-i,h}) + V_{i,h+1}^{(\widehat{m}_i^\star \diamond \pi_i)\odot \pi_{-i}, \mathcal{G}}(s_{h+1}) \Big] \\
&= V_{i,h}^{(\widehat{m}_i^\star \diamond \pi_i)\odot \pi_{-i}, \mathcal{G}}(s_h),
\end{aligned}
$$

where the second inequality if by the inductive hypothesis and the third step is by the definition of $\widehat{m}_i^\star \in \operatorname{argmax}_{m_i \in \mathcal{M}_{\mathcal{S},i}} v_i((m_i \diamond \pi_i) \odot \pi_{-i})$. □

Lemma J.3. Given an approximate POSG $\widehat{\mathcal{G}}$ that satisfies Assumption C.9 with approximate transitions and emissions being $\{\widehat{\mathbb{T}}_h, \widehat{\mathbb{O}}_h\}_{h \in [H]}$, we define the approximate decoding function $\widehat{g}$ to be

$$\widehat{g}_{i,h}(s_h \mid c_h, p_{i,h}) := \mathbb{P}^{\widehat{\mathcal{G}}}(s_h \mid c_h, p_{i,h}),$$

for each $h \in [H]$, $s_h \in \mathcal{S}$, $c_h \in \mathcal{C}_h$, $p_{i,h} \in \mathcal{P}_{i,h}$. Then it holds that for any $\pi \in \Pi_{\mathcal{S}}$,

$$
\begin{aligned}
&\max_{i \in [n], u_i \in \Pi_i, j \in [n], h \in [H]} \mathbb{P}^{u_i \times \pi_{-i}, \mathcal{G}}(s_h \neq \widehat{g}_{j,h}(c_h, p_{j,h})) \\
&\le \max_{i \in [n], u_i \in \Pi_{\mathcal{S},i}} \mathbb{E}_{u_i \times \pi_{-i}}^{\mathcal{G}}[\sum_{h \in [H]} \|\mathbb{T}_h(\cdot \mid s_h, a_h) - \widehat{\mathbb{T}}_h(\cdot \mid s_h, a_h)\|_1 + \|\mathbb{O}_h(\cdot \mid s_h) - \widehat{\mathbb{O}}_h(\cdot \mid s_h)\|_1],
\end{aligned}
$$

and

$$
\begin{aligned}
&\max_{i \in [n], m_i \in \mathcal{M}_i, j \in [n], h \in [H]} \mathbb{P}^{(m_i \diamond \pi_i)\odot \pi_{-i}, \mathcal{G}}(s_h \neq \widehat{g}_{j,h}(c_h, p_{j,h})) \\
&\le \max_{i \in [n], m_i \in \mathcal{M}_{\mathcal{S},i}} \mathbb{E}_{(m_i \diamond \pi_i)\odot \pi_{-i}}^{\mathcal{G}}[\sum_{h \in [H]} \|\mathbb{T}_h(\cdot \mid s_h, a_h) - \widehat{\mathbb{T}}_h(\cdot \mid s_h, a_h)\|_1 + \|\mathbb{O}_h(\cdot \mid s_h) - \widehat{\mathbb{O}}_h(\cdot \mid s_h)\|_1].
\end{aligned}
$$

Proof. Note for any $i \in [n]$, $u_i \in \Pi_i$, $j \in [n]$, $h \in [H]$, it holds

$$\mathbb{P}^{u_i \times \pi_{-i}, \mathcal{G}}(s_h \neq \widehat{g}_{j,h}(c_h, p_{j,h})) = \frac{1}{2} \mathbb{E}_{u_i \times \pi_{-i}}^{\mathcal{G}} \sum_{s_h} |\mathbb{P}^{\mathcal{G}}(s_h \mid c_h, p_{j,h}) - \mathbb{P}^{\widehat{\mathcal{G}}}(s_h \mid c_h, p_{j,h})|,$$

due to the condition in Definition 7.2. Meanwhile,

$$
\begin{aligned}
&\frac{1}{2} \mathbb{E}_{u_i \times \pi_{-i}}^{\mathcal{G}} \sum_{s_h} |\mathbb{P}^{\mathcal{G}}(s_h \mid c_h, p_{j,h}) - \mathbb{P}^{\widehat{\mathcal{G}}}(s_h \mid c_h, p_{j,h})| \\
&\le \sum_{s_h, c_h, p_{j,h}} |\mathbb{P}^{u_i \times \pi_{-i}, \mathcal{G}}(s_h, c_h, p_{j,h}) - \mathbb{P}^{u_i \times \pi_{-i}, \widehat{\mathcal{G}}}(s_h, c_h, p_{j,h})| \\
&\le \sum_{\overline{\tau}_h} |\mathbb{P}^{u_i \times \pi_{-i}, \mathcal{G}}(\overline{\tau}_h) - \mathbb{P}^{u_i \times \pi_{-i}, \widehat{\mathcal{G}}}(\overline{\tau}_h)| \\
&\le \sum_{\overline{\tau}_H} |\mathbb{P}^{u_i \times \pi_{-i}, \mathcal{G}}(\overline{\tau}_H) - \mathbb{P}^{u_i \times \pi_{-i}, \widehat{\mathcal{G}}}(\overline{\tau}_H)| \\
&\le \mathbb{E}_{u_i \times \pi_{-i}}^{\mathcal{G}}[\sum_{h \in [H]} \|\mathbb{T}_h(\cdot \mid s_h, a_h) - \widehat{\mathbb{T}}_h(\cdot \mid s_h, a_h)\|_1 + \|\mathbb{O}_h(\cdot \mid s_h) - \widehat{\mathbb{O}}_h(\cdot \mid s_h)\|_1],
\end{aligned}
$$

where the first inequality is by Lemma J.7, the second and third inequalities are due to the fact that TV distance does not increase after marginalization, and the last inequality is by Lemma H.7. Since π_{-i} is a fixed and fully-observable Markov policy, by Lemma H.8, we have

$$\mathbb{E}^{\mathcal{G}}_{u_i\times\pi_{-i}}\left[\sum_{h\in[H]}\|\mathbb{T}_h(\cdot\,|\,s_h,a_h)-\widehat{\mathbb{T}}_h(\cdot\,|\,s_h,a_h)\|_1+\|\mathbb{O}_h(\cdot\,|\,s_h)-\widehat{\mathbb{O}}_h(\cdot\,|\,s_h)\|_1\right]$$
$$\leq \max_{u_i\in\Pi_{\mathcal{S},i}}\mathbb{E}^{\mathcal{G}}_{u_i\times\pi_{-i}}\left[\sum_{h\in[H]}\|\mathbb{T}_h(\cdot\,|\,s_h,a_h)-\widehat{\mathbb{T}}_h(\cdot\,|\,s_h,a_h)\|_1+\|\mathbb{O}_h(\cdot\,|\,s_h)-\widehat{\mathbb{O}}_h(\cdot\,|\,s_h)\|_1\right],$$

thus proving the first result of our lemma.

For the second result of our lemma, it can be proved similarly that for any $i\in[n]$, $m_i\in\mathcal{M}_i$, $j\in[n]$, $h\in[H]$,

$$\mathbb{P}^{(m_i\diamond\pi_i)\odot\pi_{-i},\mathcal{G}}(s_h\neq\widehat{g}_{j,h}(c_h,p_{j,h}))$$
$$\leq\mathbb{E}^{\mathcal{G}}_{(m_i\diamond\pi_i)\odot\pi_{-i}}\left[\sum_{h\in[H]}\|\mathbb{T}_h(\cdot\,|\,s_h,a_h)-\widehat{\mathbb{T}}_h(\cdot\,|\,s_h,a_h)\|_1+\|\mathbb{O}_h(\cdot\,|\,s_h)-\widehat{\mathbb{O}}_h(\cdot\,|\,s_h)\|_1\right].$$

By Lemma J.2, we proved the second result. □

Theorem J.4. Fix any $\epsilon,\delta\in(0,1)$. Algorithm 5 can learn a decoding function $\widehat{g}$ such that with probability $1-\delta$

$$\max_{i\in[n],u_i\in\Pi_i,j\in[n],h\in[H]}\mathbb{P}^{u_i\times\pi_{-i},\mathcal{G}}(s_h\neq\widehat{g}_{j,h}(c_h,p_{j,h}))\leq\epsilon,$$

with total sample complexity $\widetilde{\mathcal{O}}(\frac{nS^2AHO+nS^3AH}{\epsilon^2}+\frac{S^4A^2H^5}{\epsilon})$ and computational complexity $\text{POLY}(S,A,H,O,\frac{1}{\epsilon})$.

Proof. With the help of Lemma J.3, it suffices to prove

$$\max_{i\in[n],u_i\in\Pi_{\mathcal{S},i}}\mathbb{E}^{\mathcal{G}}_{u_i\times\pi_{-i}}[\sum_{h\in[H]}\|\mathbb{T}_h(\cdot\,|\,s_h,a_h)-\widehat{\mathbb{T}}_h(\cdot\,|\,s_h,a_h)\|_1+\|\mathbb{O}_h(\cdot\,|\,s_h)-\widehat{\mathbb{O}}_h(\cdot\,|\,s_h)\|_1]\leq\epsilon.$$

The following proof procedure follows similarly to that of Theorem H.4. For each $h\in[H]$ and $s_h\in\mathcal{S}$, we define

$$p_h(s_h)=\max_{i\in[n],u_i\in\Pi_{\mathcal{S},i}}d_h^{u_i\times\pi_{-i}}(s_h).$$

Fix $\epsilon_1,\delta_1>0$, we define $\mathcal{U}(h,\epsilon_1)=\{s_h\in\mathcal{S}\,|\,p_h(s_h)\geq\epsilon_1\}$. By [32], one can learn the policy $\{\Psi_i(h,s_h)\}_{i\in[n]}$ with sample complexity $\widetilde{\mathcal{O}}(\frac{S^2A_iH^4}{\epsilon_1})$ such that $\max_{i\in[n]}d_h^{\Psi_i(h,s_h)\times\pi_{-i}}(s_h)\geq\frac{p_h(s_h)}{2}$ for each $s_h\in\mathcal{U}(h,\epsilon_1)$ with probability $1-n\cdot\delta_1$. Now we assume this event holds for any $h\in[H]$ and $s_h\in\mathcal{U}(h,\epsilon_1)$. For each $s_h\in\mathcal{S}$ and $a_h\in\mathcal{A}$, we have executed each policy $\{\Psi_i(h,s_h)\times\pi_{-i}\}_{i\in[n]}$ for the first $h-1$ steps followed by an action $a_h\in\mathcal{A}$ for N episodes and denote the total number of episodes that s_h and a_h are visited as $N_h(s_h,a_h)$, and $N_h(s_h)=\sum_{a\in\mathcal{A}}N_h(s_h,a)$. Then with probability $1-e^{-N\epsilon_1/8}$, we have $N_h(s_h,a_h)\geq\frac{Np_h(s_h)}{2}$ by Chernoff bound. Now conditioned on this event, we are ready to evaluate the following for any

$i \in [n]$, and $u_i = \Psi_i(h, s_h) \in \Pi_{\mathcal{S},i}$:

$$
\begin{aligned}
&\mathbb{E}^{\mathcal{G}}_{u_i \times \pi_{-i}} \|\mathbb{T}_h(\cdot \mid s_h, a_h) - \widehat{\mathbb{T}}_h(\cdot \mid s_h, a_h)\|_1 \\
&= \sum_{s_h, a_h} d_h^{u_i \times \pi_{-i}}(s_h)(u_i \times \pi_{-i})_h(a_h \mid s_h) \|\mathbb{T}_h(\cdot \mid s_h, a_h) - \widehat{\mathbb{T}}_h(\cdot \mid s_h, a_h)\|_1 \\
&\le 2 \cdot S\epsilon_1 + \sum_{s_h \in \mathcal{U}(h,\epsilon_1), a_h} d_h^{u_i \times \pi_{-i}}(s_h)(u_i \times \pi_{-i})_h(a_h \mid s_h) \sqrt{\frac{S \log(1/\delta_2)}{N_h(s_h, a_h)}} \\
&\le 2 \cdot S\epsilon_1 + \sum_{s_h \in \mathcal{U}(h,\epsilon_1), a_h} d_h^{u_i \times \pi_{-i}}(s_h)(u_i \times \pi_{-i})_h(a_h \mid s_h) \sqrt{\frac{2S \log(1/\delta_2)}{N p_h(s_h)}} \\
&\le 2 \cdot S\epsilon_1 + \sum_{s_h} \sqrt{d_h^{u_i \times \pi_{-i}}(s_h)} \sqrt{\frac{S \log(1/\delta_2)}{N}} \\
&\le 2 \cdot S\epsilon_1 + S\sqrt{\frac{\log(1/\delta_2)}{N}},
\end{aligned}
$$

where the second step is by Lemma J.8, and the last step is by Cauchy-Schwarz inequality. Similarly,

$$
\begin{aligned}
&\mathbb{E}^{\mathcal{P}}_{u_i \times \pi_{-i}} \|\mathbb{O}_h(\cdot \mid s_h) - \widehat{\mathbb{O}}_h(\cdot \mid s_h)\|_1 \\
&= \sum_{s_h} d_h^{u_i \times \pi_{-i}}(s_h) \|\mathbb{O}_h(\cdot \mid s_h) - \widehat{\mathbb{O}}_h(\cdot \mid s_h)\|_1 \\
&\le 2 \cdot S\epsilon_1 + \sum_{s_h \in \mathcal{U}(h,\epsilon_1)} d_h^{u_i \times \pi_{-i}}(s_h) \sqrt{\frac{O \log(1/\delta_2)}{N_h(s_h)}} \\
&\le 2 \cdot S\epsilon_1 + \sum_{s_h \in \mathcal{U}(h,\epsilon_1)} d_h^{u_i \times \pi_{-i}}(s_h) \sqrt{\frac{O \log(1/\delta_2)}{N p_h(s_h)}} \\
&\le 2 \cdot S\epsilon_1 + \sum_{s_h \in \mathcal{U}(h,\epsilon_1)} \sqrt{d_h^{u_i \times \pi_{-i}}(s_h)} \sqrt{\frac{O \log(1/\delta_2)}{N}} \\
&\le 2 \cdot S\epsilon_1 + \sqrt{\frac{SO \log(1/\delta_2)}{N}},
\end{aligned}
$$

where the second step is by Lemma J.9, and the last step is by Cauchy-Schwarz inequality. Therefore, by a union bound, all high probability events hold with probability

$$1 - SHn\delta_1 - SHAe^{-N\epsilon_1/8} - SAH\delta_2.$$

Therefore, we can choose $N = \widetilde{\mathcal{O}}(\frac{S^2+SO}{\epsilon^2})$ and $\epsilon_1 = \mathcal{O}(\frac{\epsilon}{S})$leading to the total sample complexity

$$SHA(nN + \widetilde{\mathcal{O}}(\frac{S^3AH^4}{\epsilon})) = \widetilde{\mathcal{O}}(\frac{nS^2AHO + nS^3AH}{\epsilon^2} + \frac{S^4A^2H^5}{\epsilon}).$$

□

Note that although our Algorithm 5 and Theorem J.4 are stated for NE/CCE, it can also handle CE with simple modifications, where the key observation is that the strategy modification $m_i \in \mathcal{M}_{\mathcal{S},i}$ can also be regarded as a Markov policy in an *extended* MDP marginalized by π_{-i} defined below.

Definition J.5. We define $\mathcal{M}^{\text{extended}}(\pi)$ to be an MDP for agent i, where for each $h \in [H]$, the state is $(s_h, a_{i,h})$, the action is some modified action $a'_{i,h}$, the transition is defined as $\mathbb{T}_h^{\text{extended}}(s_{h+1}, a_{i,h+1} \mid s_h, a_{i,h}, a'_{i,h}) := \mathbb{E}_{a_{-i,h} \sim \pi_h(\cdot \mid s_h, a_{i,h})}[\mathbb{T}_h(s_{h+1} \mid s_h, a'_{i,h}, a_{-i,h})\pi_{h+1}(a_{i,h+1} \mid s_{h+1})]$, where we slightly abuse the notation of $\pi_h(a_{-i,h} \mid s_h, a_{i,h})$ and $\pi_h(a_{i,h} \mid s_h)$ by defining them as the posterior and marginal distribution induced by the joint distribution $\pi_h(a_h \mid s_h)$. Similarly, the reward is given by $r_h^{\text{extended}}(s_h, a_{i,h}, a'_{i,h}) := \mathbb{E}_{a_{-i,h} \sim \pi_h(\cdot \mid s_h, a_{i,h})}[r_h(s_h, a'_{i,h}, a_{-i,h})]$.

With the help of such an extended MDP, we can develop Algorithm 6, which is a CE version of Algorithm 5 with the following guarantees.

Theorem J.6. Fix any $\epsilon, \delta \in (0,1)$. Algorithm 6 can learn a decoding function $\widehat{g}$ such that

$$\max_{i\in[n], m_i\in\mathcal{M}_i, j\in[n], h\in[H]} \mathbb{P}^{(m_i\diamond\pi_i)\odot\pi_{-i},\mathcal{G}}(s_h \neq \widehat{g}_{j,h}(c_h, p_{j,h})) \leq \epsilon,$$

with total sample complexity $\widetilde{\mathcal{O}}(\frac{nS^2A^3HO+nS^3A^4H}{\epsilon^2} + \frac{S^4A^6H^5}{\epsilon})$ and computational complexity $\mathrm{POLY}(S, A, H, O, \frac{1}{\epsilon})$.

Proof. Due to the construction of $\mathcal{G}^{\text{extended}}(\pi_{-i})$, the proof of Theorem J.4 readily applies, where the only difference is that the state space of $\mathcal{G}^{\text{extended}}(\pi_{-i})$ is SA_i, larger than that of $\mathcal{G}(\pi_{-i})$ by a factor of A_i thus proving our theorem. □

Lemma J.7. Suppose we can sample from a joint distribution $P \in \Delta(\mathcal{X}\times\mathcal{Y})$ for some finite $\mathcal{X}$, $\mathcal{Y}$ i.i.d. Then we can learn an approximate distribution $Q \in \Delta(\mathcal{X}\times\mathcal{Y})$ with sample complexity $\Theta\left(\frac{|\mathcal{X}||\mathcal{Y}|+\log 1/\delta}{\epsilon^2}\right)$ such that

$$\mathbb{E}_{x\sim P}\sum_{y\in\mathcal{Y}} |P(y\,|\,x) - Q(y\,|\,x)| \leq 2\sum_{x\in\mathcal{X}, y\in\mathcal{Y}} |P(x,y) - Q(x,y)| \leq \epsilon,$$

with probability $1-\delta$.

Proof. Note the following holds

$$\begin{aligned}\sum_{x\in\mathcal{X},y\in\mathcal{Y}} |P(x,y) - Q(x,y)| &= \sum_{x\in\mathcal{X},y\in\mathcal{Y}} |P(x,y) - P(x)Q(y\,|\,x) + P(x)Q(y\,|\,x) - Q(x,y)| \\ &\geq \sum_{x\in\mathcal{X},y\in\mathcal{Y}} |P(x,y) - P(x)Q(y\,|\,x)| - |P(x)Q(y\,|\,x) - Q(x,y)|.\end{aligned}$$

Therefore, we have

$$\begin{aligned}\mathbb{E}_{x\sim P}\sum_{y\in\mathcal{Y}} |P(y\,|\,x) - Q(y\,|\,x)| &\leq \sum_{x\in\mathcal{X},y\in\mathcal{Y}} |P(x,y) - Q(x,y)| + \sum_{x\in\mathcal{X},y\in\mathcal{Y}} |P(x) - Q(x)| \\ &\leq 2\sum_{x\in\mathcal{X},y\in\mathcal{Y}} |P(x,y) - Q(x,y)|. \end{aligned} \tag{J.2}$$

By the sample complexity of learning discrete distributions [10], we can learn Q such that $\sum_{x\in\mathcal{X},y\in\mathcal{Y}} |P(x,y) - Q(x,y)| \leq \epsilon$ in sample complexity $\Theta\left(\frac{|\mathcal{X}||\mathcal{Y}|+\log 1/\delta}{\epsilon^2}\right)$ with probability $1-\delta$. Thus, we proved our lemma. □

Lemma J.8 (Concentration on transition)**.** Fix $\delta > 0$ and dataset $\{\overline{\tau}_H^k\}_{k\in[N]}$ sampled from $\mathcal{P}$ under policy $\pi \in \Pi^{\text{gen}}$. We define for each $h \in [H]$, $(s_h, a_h, s_{h+1}) \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}$

$$\begin{aligned}N_h(s_h, a_h) &= \sum_{k\in[N]} \mathbb{1}[s_h^k = s_h, a_h^k = a_h] \\ N_h(s_h, a_h, s_{h+1}) &= \sum_{k\in[N]} \mathbb{1}[s_h^k = s_h, a_h^k = a_h, s_{h+1}^k = s_h].\end{aligned}$$

Then with probability $1-\delta$, it holds that for any $k \in [K], h \in [H], s_h \in \mathcal{S}, a_h \in \mathcal{A}$:

$$\|\mathbb{T}_h(\cdot\,|\,s_h, a_h) - \widehat{\mathbb{T}}_h(\cdot\,|\,s_h, a_h)\|_1 \leq C_1\sqrt{\frac{S\log(SAHK/\delta)}{\max\{N_h(s_h, a_h), 1\}}},$$

for some absolute constant $C_1 > 0$, where we define $\widehat{\mathbb{T}}_h(s_{h+1}\,|\,s_h, a_h) = \frac{N_h(s_h, a_h, s_{h+1})}{\max\{N_h(s_h, a_h), 1\}}$.

Proof. This is done by firstly bounding $\|\mathbb{T}_h(\cdot\,|\,s_h, a_h) - \widehat{\mathbb{T}}_h(\cdot\,|\,s_h, a_h)\|_1$ for specific k, h, s_h, a_h according to [10] and then taking union bound for all $k \in [K], h \in [H], s_h \in \mathcal{S}, a_h \in \mathcal{A}$. □

Lemma J.9 (Concentration on emission)**.** Fix $\delta > 0$ and dataset $\{\overline{\tau}_H^k\}_{k\in[N]}$ sampled from $\mathcal{P}$ under some policy $\pi \in \Pi^{\text{gen}}$. We define for each $h \in [H]$, $(s_h, o_h) \in \mathcal{S} \times \mathcal{O}$

$$N_h(s_h, o_h) = \sum_{k\in[N]} \mathbb{1}[s_h^k = s_h, o_h^k = o_h]$$
$$N_h(s_h) = \sum_{k\in[N]} \mathbb{1}[s_h^k = s_h].$$

Then with probability $1-\delta$, it holds that

$$\|\mathbb{O}_h(\cdot \mid s_h) - \widehat{\mathbb{O}}_h(\cdot \mid s_h)\|_1 \le C_2\sqrt{\frac{O\log(SHK/\delta)}{\max\{N_h^k(s_h), 1\}}},$$

for some absolute constant $C_2 > 0$, where we define $\widehat{\mathbb{O}}_h(o_h \mid s_h) = \frac{N_h(s_h,o_h)}{\max\{N_h(s_h),1\}}$.

Proof. This is done by firstly bounding $\|\mathbb{O}_h(\cdot \mid s_h) - \widehat{\mathbb{O}}_h(\cdot \mid s_h)\|_1$ for specific k, h, s_h according to [10] and then taking union bound for all $k \in [K], h \in [H], s_h \in \mathcal{S}$. □

Lemma J.10. Fix $\delta > 0$. With probability $1-\delta$, it holds that for any $k \in [K], h \in [H], s_h \in \mathcal{S}$:

$$\sum_{o_{h+1}} \left|\mathbb{P}^{\mathcal{G}}(o_{h+1} \mid s_h, a_h) - \widehat{\mathbb{J}}_h^k(o_{h+1} \mid s_h, a_h)\right| \le C_3\sqrt{\frac{O\log(SHKA/\delta)}{N_h^k(s_h, a_h)}},$$

where $\widehat{\mathbb{J}}_h^k$ is defined in Algorithm 7.

Proof. This is done by firstly bounding $\sum_{o_{h+1}} \left|\mathbb{P}^{\mathcal{G}}(o_{h+1} \mid s_h, a_h) - \widehat{\mathbb{J}}_h^k(o_{h+1} \mid s_h, a_h)\right|$ for specific k, h, s_h, a_h according to [10] and then taking union bound for all $k \in [K], h \in [H], s_h \in \mathcal{S}, a_h \in \mathcal{A}$. □

From now on, we shall use the bonus as

$$b_h^k(s_h, a_h) = \min\left\{C_3(H-h)\sqrt{\frac{O\log(SAHK/\delta)}{\max\{N_h^k(s_h, a_h), 1\}}}, 2(H-h)\right\} \tag{J.3}$$

for some absolute constant $C_3 > 0$.

Before presenting our technical analysis, we define the following notation for the ease of intermediate analysis. We define the following approximate value function for any policy $\pi \in \Pi$ in a backward way for $h \in [H]$:

$$\widehat{V}_{i,h}^{\pi,\mathcal{G}}(c_h) := \mathbb{E}_{s_h,p_h\sim\widehat{P}_h(\cdot,\cdot\mid\widehat{c}_h)}\mathbb{E}_{\omega_h,\{a_{j,h}\sim\pi_{j,h}(\cdot\mid\omega_{j,h},c_h,p_{j,h})\}_{j\in[n]}}$$
$$\mathbb{E}_{s_{h+1}\sim\mathbb{T}_h(\cdot\mid s_h,a_h),o_{h+1}\sim\mathbb{O}_{h+1}(\cdot\mid s_{h+1})}\left[r_{i,h}(s_h,a_h) + V_{i,h+1}^{\pi,\mathcal{G}}(c_{h+1})\right],$$
$$\widehat{Q}_{i,h}^{\pi,\mathcal{G}}(c_h,\gamma_h) := \mathbb{E}_{s_h,p_h\sim\widehat{P}_h(\cdot,\cdot\mid\widehat{c}_h)}\mathbb{E}_{\{a_{j,h}\sim\gamma_{j,h}(\cdot\mid p_{j,h})\}_{j\in[n]}}$$
$$\mathbb{E}_{s_{h+1}\sim\mathbb{T}_h(\cdot\mid s_h,a_h),o_{h+1}\sim\mathbb{O}_{h+1}(\cdot\mid s_{h+1})}\left[r_{i,h}(s_h,a_h) + V_{i,h+1}^{\pi,\mathcal{G}}(c_{h+1})\right],$$

for each $(i, c_h) \in [n] \times \mathcal{C}_h$, where we define $\widehat{V}_{i,H+1}^{\pi,\mathcal{G}}(c_{H+1}) = 0$.

Intuitively, this definition of $\widehat{V}_{i,h}^{\pi,\mathcal{G}}(c_h)$ mimics the Bellman equation of ground-truth value function $V_{i,h}^{\pi,\mathcal{G}}(c_h)$ by replacing the ground-truth belief $\mathbb{P}^{\mathcal{G}}(s_h, p_h \mid c_h)$ by $\widehat{P}_h(s_h, p_h \mid \widehat{c}_h)$. Next, we point out the following quantitative bound when using $\widehat{V}_{i,h}^{\pi,\mathcal{G}}(c_h)$ to approximate $V_{i,h}^{\pi,\mathcal{G}}(c_h)$.

Meanwhile, we also define the error for the belief as follows

$$\epsilon_{\text{belief}} := \max_{\pi\in\Pi} \|\mathbb{P}^{\mathcal{G}}(\cdot,\cdot \mid c_h) - \widehat{P}_h(\cdot,\cdot \mid \widehat{c}_h)\|_1.$$

Lemma J.11. For any $\pi', \pi \in \Pi$, it holds

$$\mathbb{E}^{\mathcal{G}}_{\pi'} |V^{\pi,\mathcal{G}}_{i,h}(c_h) - \widehat{V}^{\pi,\mathcal{G}}_{i,h}(c_h)| \le (H-h+1)^2 \epsilon_{\text{belief}}.$$

Proof. It follows directly by combining Lemma 4 and Lemma 8 of [43]. □

Meanwhile, note that although in Algorithm 7, the value we maintain has input $\widehat{c}_h$ instead of c_h for computational efficiency, we extend the definition of those values to also accept c_h as inputs as follows (with a slight abuse of notation):

$$\begin{aligned}
Q^{\text{high},k}_{i,h}(c_h, \gamma_h) &:= Q^{\text{high},k}_{i,h}(\widehat{c}_h, \gamma_h) \\
Q^{\text{high},k}_{i,h}(c_h, p_h, s_h, a_h) &:= Q^{\text{high},k}_{i,h}(\widehat{c}_h, p_h, s_h, a_h) \\
V^{\text{high},k}_{i,h}(c_h) &:= V^{\text{high},k}_{i,h}(\widehat{c}_h) \\
Q^{\text{high},k}_{i,h}(c_h, \gamma_h) &:= Q^{\text{high},k}_{i,h}(\widehat{c}_h, \gamma_h) \\
Q^{\text{high},k}_{i,h}(c_h, p_h, s_h, a_h) &:= Q^{\text{high},k}_{i,h}(\widehat{c}_h, p_h, s_h, a_h) \\
V^{\text{high},k}_{i,h}(c_h) &:= V^{\text{high},k}_{i,h}(\widehat{c}_h),
\end{aligned}$$

where we recall that $\widehat{c}_h = \text{Compress}_h(c_h)$.

Lemma J.12 (Optimism 1 for NE/CCE)**.** With probability $1-\delta$, for any $k \in [K]$, for Algorithm 7, it holds that for any $i \in [n]$, $\pi'_i \in \Pi_i$, $h \in [H]$

$$\begin{aligned}
Q^{\text{high},k}_{i,h}(\widehat{c}_h, \gamma_h) &\ge \widehat{Q}^{\pi'_i \times \pi^k_{-i}, \mathcal{G}}_{i,h}(c_h, \gamma_h) \\
V^{\text{high},k}_{i,h}(\widehat{c}_h) &\ge \widehat{V}^{\pi'_i \times \pi^k_{-i}, \mathcal{G}}_{i,h}(c_h),
\end{aligned}$$

where we recall that $\widehat{c}_h = \text{Compress}_h(c_h)$.

Proof. We will prove by backward induction. Obviously, it holds for $h = H+1$. Now we assume the lemma holds for $h+1$. Now we notice that by definition,

$$\begin{aligned}
&Q^{\text{high},k}_{i,h}(c_h, \gamma_h) = \mathbb{E}_{s_h, p_h \sim \widehat{P}_h(\cdot,\cdot \mid \widehat{c}_h)} \mathbb{E}_{\{a_{j,h} \sim \gamma_{j,h}(\cdot \mid p_{j,h})\}_{j \in [n]}} \left[Q^{\text{high},k}_{i,h}(c_h, p_h, s_h, a_h) \right] \\
&= \mathbb{E}_{s_h, p_h \sim \widehat{P}_h(\cdot,\cdot \mid \widehat{c}_h)} \mathbb{E}_{\{a_{j,h} \sim \gamma_{j,h}(\cdot \mid p_{j,h})\}_{j \in [n]}} \min \left\{ r_{i,h}(s_h, a_h) + b^{k-1}_h(s_h, a_h) + \mathbb{E}_{o_{h+1} \sim \widehat{\mathbb{I}}^{k-1}_h(\cdot \mid s_h, a_h)} \left[V^{\text{high},k}_{i,h+1}(c_{h+1}) \right], H-h+1 \right\} \\
&\ge \mathbb{E}_{s_h, p_h \sim \widehat{P}_h(\cdot,\cdot \mid \widehat{c}_h)} \mathbb{E}_{\{a_{j,h} \sim \gamma_{j,h}(\cdot \mid p_{j,h})\}_{j \in [n]}} \\
&\qquad \min \left\{ r_{i,h}(s_h, a_h) + b^{k-1}_h(s_h, a_h) + \mathbb{E}_{o_{h+1} \sim \widehat{\mathbb{I}}^{k-1}_h(\cdot \mid s_h, a_h)} \left[\widehat{V}^{\pi'_i \times \pi^k_{-i}, \mathcal{G}}_{i,h+1}(c_{h+1}) \right], H-h+1 \right\},
\end{aligned}$$

where the last step is by inductive hypothesis. Now note that for any s_h, p_h, a_h, we have

$$\begin{aligned}
&b^{k-1}_h(s_h, a_h) + \mathbb{E}_{o_{h+1} \sim \widehat{\mathbb{I}}^{k-1}_h(\cdot \mid s_h, a_h)} \left[\widehat{V}^{\pi'_i \times \pi^k_{-i}, \mathcal{G}}_{i,h+1}(c_{h+1}) \right] \\
&\ge b^{k-1}_h(s_h, a_h) - (H-h) \| \widehat{\mathbb{I}}^{k-1}_h(\cdot \mid s_h, a_h) - \mathbb{P}^{\mathcal{G}}(\cdot \mid s_h, a_h) \|_1 + \mathbb{E}_{s_{h+1} \sim \mathbb{T}_h(\cdot \mid s_h, a_h), o_{h+1} \sim \mathbb{O}_{h+1}(\cdot \mid s_{h+1})} \left[\widehat{V}^{\pi'_i \times \pi^k_{-i}, \mathcal{G}}_{i,h+1}(c_{h+1}) \right] \\
&\ge \mathbb{E}_{s_{h+1} \sim \mathbb{T}_h(\cdot \mid s_h, a_h), o_{h+1} \sim \mathbb{O}_{h+1}(\cdot \mid s_{h+1})} \left[\widehat{V}^{\pi'_i \times \pi^k_{-i}, \mathcal{G}}_{i,h+1}(c_{h+1}) \right],
\end{aligned}$$

where we notice $\mathbb{P}^{\mathcal{G}}(o_{h+1} \mid s_h, a_h) = \sum_{s_{h+1}} \mathbb{O}_{h+1}(o_{h+1} \mid s_{h+1}) \mathbb{T}_h(s_{h+1} \mid s_h, a_h)$ for the first inequality, and the second inequality comes from the construction of our bonus $b^{k-1}_h(s_h, a_h)$ in Equation (J.3) and Lemma J.10. Meanwhile, by the definition of value functions, it holds that $\mathbb{E}_{s_{h+1} \sim \mathbb{T}_h(\cdot \mid s_h, a_h), o_{h+1} \sim \mathbb{O}_{h+1}(\cdot \mid s_{h+1})} \left[\widehat{V}^{\pi'_i \times \pi^k_{-i}, \mathcal{G}}_{i,h+1}(c_{h+1}) \right] \le H-h$. Therefore, we have

$$\begin{aligned}
&\min \left\{ r_{i,h}(s_h, a_h) + b^{k-1}_h(s_h, a_h) + \mathbb{E}_{s_{h+1}, o_{h+1} \sim \widehat{\mathbb{I}}^{k-1}_h(\cdot,\cdot \mid s_h, a_h)} \left[V^{\pi'_i \times \pi^k_{-i}, \mathcal{G}}_{i,h+1}(c_{h+1}) \right], H-h+1 \right\} \\
&\quad \ge r_{i,h}(s_h, a_h) + \mathbb{E}_{s_{h+1} \sim \mathbb{T}_h(\cdot \mid s_h, a_h), o_{h+1} \sim \mathbb{O}_{h+1}(\cdot \mid s_{h+1})} \left[\widehat{V}^{\pi'_i \times \pi^k_{-i}, \mathcal{G}}_{i,h+1}(c_{h+1}) \right].
\end{aligned}$$

Now we conclude

$$
\begin{aligned}
&Q_{i,h}^{\text{high},k}(c_h, \gamma_h) \\
&\geq \mathbb{E}_{s_h, p_h \sim \widehat{P}_h(\cdot,\cdot \mid \widehat{c}_h)} \mathbb{E}_{\{a_{j,h} \sim \gamma_{j,h}(\cdot \mid p_{j,h})\}_{j\in[n]}} \mathbb{E}_{s_{h+1}\sim \mathbb{T}_h(\cdot \mid s_h, a_h), o_{h+1}\sim \mathbb{O}_{h+1}(\cdot \mid s_{h+1})} \left[r_{i,h}(s_h, a_h) + \widehat{V}_{i,h+1}^{\pi_i' \times \pi_{-i}^k, \mathcal{G}}(c_{h+1}) \right] \\
&= \widehat{Q}_{i,h}^{\pi_i' \times \pi_{-i}^k, \mathcal{G}}(c_h, \gamma_h).
\end{aligned}
$$

By definition, we have $Q_{i,h}^{\text{high},k}(c_h, \gamma_h) = Q_{i,h}^{\text{high},k}(\widehat{c}_h, \gamma_h)$, thus proving $Q_{i,h}^{\text{high},k}(\widehat{c}_h, \gamma_h) \geq \widehat{Q}_{i,h}^{\pi_i' \times \pi_{-i}^k, \mathcal{G}}(c_h, \gamma_h)$. Now for the value function, note that

$$
\begin{aligned}
V_{i,h}^{\text{high},k}(c_h) &= \mathbb{E}_{\omega_h} Q_{i,h}^{\text{high},k}(c_h, \{\pi_{j,h}^k(\cdot \mid \omega_{j,h}, \widehat{c}_h, \cdot)\}_{j\in[n]}) \\
&\geq \mathbb{E}_{\omega_h'} \mathbb{E}_{\omega_h} Q_{i,h}^{\text{high},k}(c_h, \{\pi_{j,h}^k(\cdot \mid \omega_{j,h}, \widehat{c}_h, \cdot)\}_{j\in[n]\setminus\{i\}}, \pi_{i,h}'(\cdot \mid \omega_{i,h}', c_h, \cdot)) \\
&\geq \mathbb{E}_{\omega_h'} \mathbb{E}_{\omega_h} \widehat{Q}_{i,h}^{\pi_i' \times \pi_{-i}^k, \mathcal{G}}(c_h, \{\pi_{j,h}^k(\cdot \mid \omega_{j,h}, \widehat{c}_h, \cdot)\}_{j\in[n]\setminus\{i\}}, \pi_{i,h}'(\cdot \mid \omega_{i,h}', c_h, \cdot)) \\
&= \widehat{V}_{i,h}^{\pi_i' \times \pi_{-i}^k, \mathcal{G}}(c_h),
\end{aligned}
$$

where the first step is by the property of Bayesian CCE, and the second step is by $Q_{i,h}^{\text{high},k}(c_h, \gamma_h) \geq \widehat{Q}_{i,h}^{\pi_i' \times \pi_{-i}^k, \mathcal{G}}(c_h, \gamma_h)$ for any $\gamma_h \in \Gamma_h$. Again by definition, we proved $V_{i,h}^{\text{high},k}(\widehat{c}_h) = V_{i,h}^{\text{high},k}(c_h) \geq \widehat{V}_{i,h}^{\pi_i' \times \pi_{-i}^k, \mathcal{G}}(c_h)$. □

Lemma J.13 (Optimism 1 for CE)**.** With probability $1-\delta$, for any $k \in [K]$, for Algorithm 7, it holds that for any $i \in [n]$, $m_i \in \mathcal{M}_i$, $h \in [H]$

$$
\begin{aligned}
Q_{i,h}^{\text{high},k}(\widehat{c}_h, \gamma_h) &\geq \widehat{Q}_{i,h}^{(m_i \diamond \pi_i^k) \odot \pi_{-i}^k, \mathcal{G}}(c_h, \gamma_h) \\
V_{i,h}^{\text{high},k}(\widehat{c}_h) &\geq \widehat{V}_{i,h}^{(m_i \diamond \pi_i^k) \odot \pi_{-i}^k, \mathcal{G}}(c_h).
\end{aligned}
$$

Proof. We will prove by backward induction. Obviously, it holds for $h = H+1$. Now we assume the lemma holds for $h+1$. Now we notice that by definition,

$$
\begin{aligned}
&Q_{i,h}^{\text{high},k}(c_h, \gamma_h) = \mathbb{E}_{s_h, p_h \sim \widehat{P}_h(\cdot,\cdot \mid \widehat{c}_h)} \mathbb{E}_{\{a_{j,h} \sim \gamma_{j,h}(\cdot \mid p_{j,h})\}_{j\in[n]}} \left[Q_{i,h}^{\text{high},k}(c_h, p_h, s_h, a_h) \right] \\
&= \mathbb{E}_{s_h, p_h \sim \widehat{P}_h(\cdot,\cdot \mid \widehat{c}_h)} \mathbb{E}_{\{a_{j,h} \sim \gamma_{j,h}(\cdot \mid p_{j,h})\}_{j\in[n]}} \min \left\{ r_{i,h}(s_h, a_h) + b_h^{k-1}(s_h, a_h) + \mathbb{E}_{o_{h+1} \sim \widehat{\mathbb{J}}_h^{k-1}(\cdot \mid s_h, a_h)} \left[V_{i,h+1}^{\text{high},k}(c_{h+1}) \right], H - h + 1 \right\} \\
&\geq \mathbb{E}_{s_h, p_h \sim \widehat{P}_h(\cdot,\cdot \mid \widehat{c}_h)} \mathbb{E}_{\{a_{j,h} \sim \gamma_{j,h}(\cdot \mid p_{j,h})\}_{j\in[n]}} \\
&\qquad \min \left\{ r_{i,h}(s_h, a_h) + b_h^{k-1}(s_h, a_h) + \mathbb{E}_{o_{h+1} \sim \widehat{\mathbb{J}}_h^{k-1}(\cdot \mid s_h, a_h)} \left[\widehat{V}_{i,h+1}^{(m_i \diamond \pi_i^k) \odot \pi_{-i}^k, \mathcal{G}}(c_{h+1}) \right], H - h + 1 \right\},
\end{aligned}
$$

where the last step is by inductive hypothesis. Now note that for any s_h, p_h, a_h, we have

$$
\begin{aligned}
&b_h^{k-1}(s_h, a_h) + \mathbb{E}_{o_{h+1} \sim \widehat{\mathbb{J}}_h^{k-1}(\cdot \mid s_h, a_h)} \left[\widehat{V}_{i,h+1}^{(m_i \diamond \pi_i^k) \odot \pi_{-i}^k, \mathcal{G}}(c_{h+1}) \right] \\
&\geq b_h^{k-1}(s_h, a_h) - (H-h) \| \widehat{\mathbb{J}}_h^{k-1}(\cdot \mid s_h, a_h) - \mathbb{P}^{\mathcal{G}}(\cdot \mid s_h, a_h) \|_1 + \mathbb{E}_{s_{h+1} \sim \mathbb{T}_h(\cdot \mid s_h, a_h), o_{h+1} \sim \mathbb{O}_{h+1}(\cdot \mid s_{h+1})} \left[\widehat{V}_{i,h+1}^{(m_i \diamond \pi_i^k) \odot \pi_{-i}^k, \mathcal{G}}(c_{h+1}) \right] \\
&\geq \mathbb{E}_{s_{h+1} \sim \mathbb{T}_h(\cdot \mid s_h, a_h), o_{h+1} \sim \mathbb{O}_{h+1}(\cdot \mid s_{h+1})} \left[\widehat{V}_{i,h+1}^{(m_i \diamond \pi_i^k) \odot \pi_{-i}^k, \mathcal{G}}(c_{h+1}) \right],
\end{aligned}
$$

where we notice $\mathbb{P}^{\mathcal{G}}(o_{h+1} \mid s_h, a_h) = \sum_{s_{h+1}} \mathbb{O}_{h+1}(o_{h+1} \mid s_{h+1}) \mathbb{T}_h(s_{h+1} \mid s_h, a_h)$ for the first inequality, and the second inequality comes from the construction of our bonus $b_h^{k-1}(s_h, a_h)$ in Equation (J.3) and Lemma J.10. Meanwhile, by the definition of value functions, it holds that $\mathbb{E}_{s_{h+1} \sim \mathbb{T}_h(\cdot \mid s_h, a_h), o_{h+1} \sim \mathbb{O}_{h+1}(\cdot \mid s_{h+1})} \left[\widehat{V}_{i,h+1}^{(m_i \diamond \pi_i^k) \odot \pi_{-i}^k, \mathcal{G}}(c_{h+1}) \right] \leq H - h$. Therefore, we have

$$
\begin{aligned}
&\min \left\{ r_{i,h}(s_h, a_h) + b_h^{k-1}(s_h, a_h) + \mathbb{E}_{s_{h+1}, o_{h+1} \sim \widehat{\mathbb{J}}_h^{k-1}(\cdot,\cdot \mid s_h, a_h)} \left[V_{i,h+1}^{(m_i \diamond \pi_i^k) \odot \pi_{-i}^k, \mathcal{G}}(c_{h+1}) \right], H - h + 1 \right\} \\
&\quad \geq r_{i,h}(s_h, a_h) + \mathbb{E}_{s_{h+1} \sim \mathbb{T}_h(\cdot \mid s_h, a_h), o_{h+1} \sim \mathbb{O}_{h+1}(\cdot \mid s_{h+1})} \left[\widehat{V}_{i,h+1}^{(m_i \diamond \pi_i^k) \odot \pi_{-i}^k, \mathcal{G}}(c_{h+1}) \right].
\end{aligned}
$$

Now we conclude

$$
\begin{aligned}
&Q_{i,h}^{\text{high},k}(c_h,\gamma_h)\\
&\geq \mathbb{E}_{s_h,p_h\sim\widehat{P}_h(\cdot,\cdot\,|\,\widehat{c}_h)}\mathbb{E}_{\{a_{j,h}\sim\gamma_{j,h}(\cdot\,|\,p_{j,h})\}_{j\in[n]}}\mathbb{E}_{s_{h+1}\sim\mathbb{T}_h(\cdot\,|\,s_h,a_h),o_{h+1}\sim\mathbb{O}_{h+1}(\cdot\,|\,s_{h+1})}\left[r_{i,h}(s_h,a_h)+\widehat{V}_{i,h+1}^{(m_i\diamond\pi_i^k)\odot\pi_{-i}^k,\mathcal{G}}(c_{h+1})\right]\\
&=\widehat{Q}_{i,h}^{(m_i\diamond\pi_i^k)\odot\pi_{-i}^k,\mathcal{G}}(c_h,\gamma_h).
\end{aligned}
$$

By definition, we have $Q_{i,h}^{\text{high},k}(c_h,\gamma_h) = Q_{i,h}^{\text{high},k}(\widehat{c}_h,\gamma_h)$, thus proving $Q_{i,h}^{\text{high},k}(\widehat{c}_h,\gamma_h) \geq \widehat{Q}_{i,h}^{(m_i\diamond\pi_i^k)\odot\pi_{-i}^k,\mathcal{G}}(c_h,\gamma_h)$. Now for the value function, note that

$$
\begin{aligned}
V_{i,h}^{\text{high},k}(c_h)&=\mathbb{E}_{\omega_h}Q_{i,h}^{\text{high},k}(c_h,\{\pi_{j,h}^k(\cdot\,|\,\omega_{j,h},\widehat{c}_h,\cdot)\}_{j\in[n]})\\
&\geq \mathbb{E}_{\omega_h}Q_{i,h}^{\text{high},k}(c_h,\{\pi_{j,h}^k(\cdot\,|\,\omega_{j,h},\widehat{c}_h,\cdot)\}_{j\in[n]\setminus\{i\}},(m_{i,h}\diamond\pi_{i,h}^k)(\cdot\,|\,\omega_{i,h},\widehat{c}_h,\cdot))\\
&\geq \mathbb{E}_{\omega_h}\widehat{Q}_{i,h}^{(m_i\diamond\pi_i^k)\odot\pi_{-i}^k,\mathcal{G}}(c_h,\{\pi_{j,h}^k(\cdot\,|\,\omega_{j,h},\widehat{c}_h,\cdot)\}_{j\in[n]\setminus\{i\}},(m_{i,h}\diamond\pi_{i,h}^k)(\cdot\,|\,\omega_{i,h},\widehat{c}_h,\cdot))\\
&=\widehat{V}_{i,h}^{(m_i\diamond\pi_i^k)\odot\pi_{-i}^k,\mathcal{G}}(c_h),
\end{aligned}
$$

where the first step is by the property of Bayesian CE, and the second step is by $Q_{i,h}^{\text{high},k}(c_h,\gamma_h) \geq \widehat{Q}_{i,h}^{(m_i\diamond\pi_i^k)\odot\pi_{-i}^k,\mathcal{G}}(c_h,\gamma_h)$ for any $\gamma_h \in \Gamma_h$. Again by definition, we proved $V_{i,h}^{\text{high},k}(\widehat{c}_h) = V_{i,h}^{\text{high},k}(c_h) \geq \widehat{V}_{i,h}^{(m_i\diamond\pi_i^k)\odot\pi_{-i}^k,\mathcal{G}}(c_h)$. □

Lemma J.14 (Pessimism)**.** With probability $1-\delta$, for any $k\in[K]$, for Algorithm 7, it holds that for any $i\in[n]$, $h\in[H]$

$$
\begin{aligned}
Q_{i,h}^{\text{low},k}(\widehat{c}_h,\gamma_h)&\leq\widehat{Q}_{i,h}^{\pi^k,\mathcal{G}}(c_h,\gamma_h)\\
V_{i,h}^{\text{low},k}(\widehat{c}_h)&\leq\widehat{V}_{i,h}^{\pi^k,\mathcal{G}}(c_h).
\end{aligned}
$$

Proof. We prove by backward induction on h. Obviously, the lemma holds for $h=H+1$. Now we assume the lemma holds for $h+1$. Similar to the proof of the previous lemma, we note by inductive hypothesis

$$
\begin{aligned}
Q_{i,h}^{\text{low},k}(c_h,\gamma_h)\leq\mathbb{E}_{s_h,p_h\sim\widehat{P}_h(\cdot,\cdot\,|\,\widehat{c}_h)}&\mathbb{E}_{\{a_{j,h}\sim\gamma_{j,h}(\cdot\,|\,p_{j,h})\}_{j\in[n]}}\\
&\max\left\{r_{i,h}(s_h,a_h)-b_h^{k-1}(s_h,a_h)+\mathbb{E}_{s_{h+1},o_{h+1}\sim\widehat{\mathbb{J}}_h^{k-1}(\cdot,\cdot\,|\,s_h,a_h)}\left[\widehat{V}_{i,h+1}^{\pi^k,\mathcal{G}}(c_{h+1})\right],0\right\},
\end{aligned}
$$

where for any s_h, p_h, a_h, we have

$$
\begin{aligned}
&-b_h^{k-1}(s_h,a_h)+\mathbb{E}_{o_{h+1}\sim\widehat{\mathbb{J}}_h^{k-1}(\cdot\,|\,s_h,a_h)}\left[V_{i,h+1}^{\text{low},k}(c_{h+1})\right]\\
&\leq -b_h^{k-1}(s_h,a_h)+(H-h)\|\widehat{\mathbb{J}}_h^{k-1}(\cdot\,|\,s_h,a_h)-\mathbb{P}^{\mathcal{G}}(\cdot\,|\,s_h,a_h)\|_1+\mathbb{E}_{s_{h+1}\sim\mathbb{T}_h(\cdot\,|\,s_h,a_h),o_{h+1}\sim\mathbb{O}_{h+1}(\cdot\,|\,s_{h+1})}\left[\widehat{V}_{i,h+1}^{\pi^k,\mathcal{G}}(c_{h+1})\right]\\
&\leq\mathbb{E}_{s_{h+1}\sim\mathbb{T}_h(\cdot\,|\,s_h,a_h),o_{h+1}\sim\mathbb{O}_{h+1}(\cdot\,|\,s_{h+1})}\left[\widehat{V}_{i,h+1}^{\pi^k,\mathcal{G}}(c_{h+1})\right],
\end{aligned}
$$

where the last step again comes from the construction of our bonus in Equation (J.3) and Lemma J.10. Therefore, we conclude

$$
\begin{aligned}
Q_{i,h}^{\text{low},k}(c_h,\gamma_h)&\leq\mathbb{E}_{s_h,p_h\sim\widehat{P}_h(\cdot,\cdot\,|\,\widehat{c}_h)}\mathbb{E}_{\{a_{j,h}\sim\gamma_{j,h}(\cdot\,|\,p_{j,h})\}_{j\in[n]}}\mathbb{E}_{o_{h+1}\sim\widehat{\mathbb{J}}_h^{k-1}(\cdot\,|\,s_h,a_h)}\left[r_{i,h}(s_h,a_h)+\widehat{V}_{i,h+1}^{\pi^k,\mathcal{G}}(c_{h+1})\right]\\
&=\widehat{Q}_{i,h}^{\pi^k,\mathcal{G}}(c_h,\gamma_h).
\end{aligned}
$$

Similarly, for value function, it holds that

$$
V_{i,h}^{\text{low},k}(c_h)=\mathbb{E}_{\omega_h}Q_{i,h}^{\text{low},k}(c_h,\{\pi_{j,h}^k(\cdot\,|\,\omega_{j,h},\widehat{c}_h,\cdot)\}_{j\in[n]})\leq\mathbb{E}_{\omega_h}\widehat{Q}_{i,h}^{\pi^k,\mathcal{G}}(c_h,\{\pi_{j,h}^k(\cdot\,|\,\omega_{j,h},\widehat{c}_h,\cdot)\}_{j\in[n]})=\widehat{V}_{i,h}^{\pi^k,\mathcal{G}}(c_h),
$$

thus proving our lemma. □

Theorem J.15 (NE/CCE version)**.** With probability $1-\delta$, Algorithm 7 enjoys the regret guarantee of

$$\sum_{k\in[K]}\max_{i\in[n]}\left(\max_{\pi_i'\in\Pi_i}V_{i,1}^{\pi_i'\times\pi_{-i}^k,\mathcal{G}}(c_1^k)-V_{i,1}^{\pi^k,\mathcal{G}}(c_1^k)\right)\le\mathcal{O}(KH^2\epsilon_{\text{belief}}+H^2\sqrt{SAOK\log(SAHK/\delta)}+H^2SA\sqrt{O\log(SAHK/\delta)}).$$

Correspondingly, this implies one can learn an $(\epsilon+H^2\epsilon_{\text{belief}})$-NE if $\mathcal{G}$ is zero-sum and $(\epsilon+H\epsilon_{\text{belief}}^{\epsilon})$-CCE if $\mathcal{G}$ is general-sum with sample complexity $\mathcal{O}(\frac{H^4SAO\log(SAHO/\delta)}{\epsilon^2})$ and computation complexity $\text{POLY}(S,A,O,H,\frac{1}{\epsilon},\log\frac{1}{\delta})$.

Proof. Note for any given $i\in[n]$ and $\pi_i'\in\Pi_i$, by Lemma J.12 and Lemma J.14, it holds

$$\max_{\pi_i'\in\Pi_i}V_{i,1}^{\pi_i'\times\pi_{-i}^k,\mathcal{G}}(c_1^k)-V_{i,1}^{\pi^k,\mathcal{G}}(c_1^k)\le V_{i,h}^{\text{high},k}(c_h^k)-V_{i,h}^{\text{low},k}(c_h^k).$$

Therefore, it suffices to bound $V_{i,h}^{\text{high},k}(c_h^k)-V_{i,h}^{\text{low},k}(c_h^k)$.

$$\begin{aligned}
&V_{i,h}^{\text{high},k}(c_h^k)-V_{i,h}^{\text{low},k}(c_h^k)\\
&=\mathbb{E}_{s_h,p_h\sim\widehat{P}_h(\cdot,\cdot\mid\widehat{c}_h^k)}\mathbb{E}_{\omega_h}\left[Q_{i,h}^{\text{high},k}(c_h^k,\{\pi_{j,h}^k(\cdot\mid\cdot,\widehat{c}_h^k,\cdot)\}_{j\in[n]})-Q_{i,h}^{\text{low},k}(c_h^k,\{\pi_{j,h}^k(\cdot\mid\cdot,\widehat{c}_h^k,\cdot)\}_{j\in[n]})\right]\\
&\le\mathbb{E}_{s_h,p_h\sim\mathbb{P}^{\mathcal{G}}(\cdot,\cdot\mid c_h^k)}\mathbb{E}_{\omega_h}\left[Q_{i,h}^{\text{high},k}(c_h^k,\{\pi_{j,h}^k(\cdot\mid\cdot,\widehat{c}_h^k,\cdot)\}_{j\in[n]})-Q_{i,h}^{\text{low},k}(c_h^k,\{\pi_{j,h}^k(\cdot\mid\cdot,\widehat{c}_h^k,\cdot)\}_{j\in[n]})\right]+(H-h+1)\epsilon_h(c_h^k)\\
&\le\mathbb{E}_{s_h,p_h\sim\mathbb{P}^{\mathcal{G}}(\cdot,\cdot\mid c_h^k)}\mathbb{E}_{\omega_h}\mathbb{E}_{\{a_{j,h}\sim\pi_{j,h}^k(\cdot\mid\omega_{j,h},\widehat{c}_h^k,p_{j,h})\}_{j\in[n]}}\left[Q_{i,h}^{\text{high},k}(c_h^k,p_h,s_h,a_h)-Q_{i,h}^{\text{low},k}(c_h^k,p_h,s_h,a_h)\right]+(H-h+1)\epsilon_h(c_h^k)\\
&\le Z_{k,h}^1+Q_{i,h}^{\text{high},k}(c_h^k,p_h^k,s_h^k,a_h^k)-Q_{i,h}^{\text{low},k}(c_h^k,p_h^k,s_h^k,a_h^k)+(H-h+1)\epsilon_h(c_h^k)\\
&\le Z_{k,h}^1+\mathbb{E}_{o_{h+1}\sim\widehat{\mathbb{P}}_h^{k-1}(\cdot\mid s_h^k,a_h^k)}\left[V_{i,h+1}^{\text{high},k}(c_{h+1})-V_{i,h+1}^{\text{low},k}(c_{h+1})\right]+(H-h+1)\epsilon_h(c_h^k)+2b_h^{k-1}(s_h^k,a_h^k)\\
&\le Z_{k,h}^1+\mathbb{E}_{o_{h+1}\sim\mathbb{P}^{\mathcal{G}}(\cdot\mid s_h^k,a_h^k)}\left[V_{i,h+1}^{\text{high},k}(c_{h+1})-V_{i,h+1}^{\text{low},k}(c_{h+1})\right]+(H-h+1)\epsilon_h(c_h^k)+3b_h^{k-1}(s_h^k,a_h^k)\\
&\le Z_{k,h}^1+Z_{k,h}^2+V_{i,h+1}^{\text{high},k}(c_{h+1}^k)-V_{i,h+1}^{\text{low},k}(c_{h+1}^k)+(H-h+1)\epsilon_h(c_h^k)+3b_h^{k-1}(s_h^k,a_h^k),
\end{aligned}$$

where we define the Martingale difference sequence as follows

$$\begin{aligned}
Z_{k,h}^1:=&\,\mathbb{E}_{s_h,p_h\sim\mathbb{P}^{\mathcal{G}}(\cdot,\cdot\mid c_h^k)}\mathbb{E}_{\omega_h}\mathbb{E}_{\{a_{j,h}\sim\pi_{j,h}^k(\cdot\mid\omega_{j,h},\widehat{c}_h^k,p_{j,h})\}_{j\in[n]}}\left[Q_{i,h}^{\text{high},k}(c_h^k,p_h,s_h,a_h)-Q_{i,h}^{\text{low},k}(c_h^k,p_h,s_h,a_h)\right]\\
&-\left(Q_{i,h}^{\text{high},k}(c_h^k,p_h^k,s_h^k,a_h^k)-Q_{i,h}^{\text{low},k}(c_h^k,p_h^k,s_h^k,a_h^k)\right)\\
Z_{k,h}^2:=&\,\mathbb{E}_{o_{h+1}\sim\mathbb{P}^{\mathcal{G}}(\cdot\mid s_h^k,a_h^k)}\left[V_{i,h+1}^{\text{high},k}(c_{h+1})-V_{i,h+1}^{\text{low},k}(c_{h+1})\right]-\left(V_{i,h+1}^{\text{high},k}(c_{h+1}^k)-V_{i,h+1}^{\text{low},k}(c_{h+1}^k)\right),
\end{aligned}$$

and the error of the belief is defined as

$$\epsilon_h(c_h^k):=\|\widehat{P}_h(\cdot,\cdot\mid\widehat{c}_h^k)-\mathbb{P}^{\mathcal{G}}(\cdot,\cdot\mid c_h^k)\|_1.$$

Since $|Z_{k,h}^1|\le H$, $|Z_{k,h}^2|\le H$, and $\epsilon_h(c_h^k)\le 2$, by Azuma-Hoeffding bound, we conclude

$$\begin{aligned}
&\sum_{k,h}Z_{k,h}^1\le\mathcal{O}(H\sqrt{HK}),\\
&\sum_{k,h}Z_{k,h}^2\le\mathcal{O}(H\sqrt{HK}),\\
&\sum_{k,h}\epsilon_h(c_h^k)\le\sum_k\mathbb{E}_{\pi^k}^{\mathcal{G}}\left[\sum_h\epsilon_h(c_h)\right]+\mathcal{O}(\sqrt{HK})\le KH\epsilon_{\text{belief}}+\mathcal{O}(\sqrt{HK}).
\end{aligned}$$

Meanwhile, by the pigeonhole principle, it holds that

$$\begin{aligned}
\sum_{k,h}b_h^{k-1}(s_h^k,a_h^k)&\le H\sqrt{O\log(SAHK/\delta)}\sum_{k,h}\frac{1}{\sqrt{\max\{1,N_h^{k-1}(s_h^k,a_h^k)\}}}\\
&\le\mathcal{O}\left(H\sqrt{O\log(SAHK/\delta)}(H\sqrt{SAK}+HSA)\right).
\end{aligned}$$

Now by Lemma J.12 and Lemma J.14 and putting everything together, we conclude

$$\sum_{k\in[K]}\max_{i\in[n]}\left(\max_{\pi_i'\in\Pi_i}\widehat{V}_{i,1}^{\pi_i'\times\pi_{-i},\mathcal{G}}(c_1^k)-\widehat{V}_{i,1}^{\pi,\mathcal{G}}(c_1^k)\right)\le KH^2\epsilon_{\text{belief}}+\mathcal{O}(H^2\sqrt{SAOK\log(SAHK/\delta)}+H^2SA\sqrt{O\log(SAHK/\delta)}),$$

Now by Lemma J.11, we proved the regret guarantees as follows

$$\sum_{k\in[K]}\max_{i\in[n]}\left(\max_{\pi_i'\in\Pi_i}V_{i,1}^{\pi_i'\times\pi_{-i}^k,\mathcal{G}}(c_1^k)-V_{i,1}^{\pi^k,\mathcal{G}}(c_1^k)\right)\le\mathcal{O}(KH^2\epsilon_{\text{belief}}+H^2\sqrt{SAOK\log(SAHK/\delta)}+H^2SA\sqrt{O\log(SAHK/\delta)}).$$

For the PAC guarantees, since we define $k^\star\in\arg\min_{k\in[K]}V_{i,1}^{\text{high},k}(c_1^k)-V_{i,1}^{\text{low},k}(c_1^k)$, we have

$$\begin{aligned}\text{CCE-gap}(\pi^{k^\star})&\le\mathcal{O}(H^2\epsilon_{\text{belief}})+V_{i,h}^{\text{high},k^\star}(c_1^{k^\star})-V_{i,h}^{\text{low},k^\star}(c_1^{k^\star})\\&\le\mathcal{O}(H^2\epsilon_{\text{belief}})+\frac{1}{K}\sum_{k\in[K]}V_{i,h}^{\text{high},k}(c_1^k)-V_{i,h}^{\text{low},k}(c_1^k)\\&\le\mathcal{O}(H^2\epsilon_{\text{belief}}+H^2\sqrt{SAO\log(SAHK/\delta)/K}+\frac{H^2SA}{K}\sqrt{O\log(SAHK/\delta)}).\end{aligned}$$

Finally, for two-player zero-sum games, we denote $\widehat{\pi}^{k^\star}$ to be the marginalized policy of $\pi^{k^\star}$. Then we have

$$\text{NE-gap}(\widehat{\pi}^{k^\star})\le\text{CCE-gap}(\pi^{k^\star}),$$

thus concluding our theorem. □

Theorem J.16 (CE version)**.** With probability $1-\delta$, Algorithm 7 enjoys the regret guarantee of

$$\sum_{k\in[K]}\max_{i\in[n]}\left(\max_{m_i'\in\mathcal{M}_i}V_{i,1}^{m_i\diamond\pi_i^k\odot\pi_{-i}^k,\mathcal{G}}(c_1^k)-V_{i,1}^{\pi^k,\mathcal{G}}(c_1^k)\right)\le\mathcal{O}(KH^2\epsilon_{\text{belief}}+H^2\sqrt{SAOK\log(SAHK/\delta)}+H^2SA\sqrt{O\log(SAHK/\delta)})$$

Correspondingly, this implies one can learn an $(\epsilon+H^2\epsilon_{\text{belief}})$-CE with sample complexity $\mathcal{O}(\frac{H^4SAO\log(SAHO/\delta)}{\epsilon^2})$

Proof. Then proof follows as that of Theorem J.15, where we only need to change the first step of the proof as

$$\widehat{V}_{i,h}^{\pi_i'\times\pi_{-i}^k,\mathcal{G}}(c_h^k)-\widehat{V}_{i,h}^{\pi^k,\mathcal{G}}(c_h^k)\le V_{i,h}^{\text{high},k}(c_h^k)-V_{i,h}^{\text{low},k}(c_h^k),$$

by Lemma J.13 and Lemma J.14, and the remaining steps are exactly the same. □

Lemma J.17 (Adapted from Theorem H.4)**.** Algorithm 8 can learn the approximate POMDP with transition $\widehat{\mathbb{T}}_{1:H}$ and emission $\widehat{\mathbb{O}}_{1:H}$ such that for any policy $\pi\in\Pi^{\text{gen}}$ and $h\in[H]$

$$\mathbb{E}_\pi^{\mathcal{G}}\left[\|\mathbb{T}_h(\cdot\,|\,s_h,a_h)-\widehat{\mathbb{T}}_h(\cdot\,|\,s_h,a_h)\|_1+\|\mathbb{O}_h(\cdot\,|\,s_h)-\widehat{\mathbb{O}}_h(\cdot\,|\,s_h)\|_1\right]\le\epsilon,$$

using sample complexity $\widetilde{\mathcal{O}}(\frac{S^2AHO+S^3AH}{\epsilon^2}+\frac{S^4A^2H^5}{\epsilon})$ with probability $1-\delta$.

Proof. Note that Algorithm 8 is essentially treating the POSG $\mathcal{G}$ as a centralized MDP and running Algorithm 4, where the only modifications we make in Algorithm 8 is that we take the controller set into considerations when learning the models. Specifically, for the transition $\widehat{\mathbb{T}}_h$, what we estimate is only $\widehat{\mathbb{T}}_h(s_{h+1}\,|\,s_h,a_{\mathcal{I}_h,h})$ instead of $\widehat{\mathbb{T}}_h(s_{h+1}\,|\,s_h,a_h)$. Therefore, the sample complexity of Algorithm 8 will not be worse than that of Algorithm 4. □

Proof of Theorem 7.7:

Note that the proof idea essentially resembles that of Theorem H.5, where we construct the model $\mathcal{G}^{\text{trunc}}$ for $\mathcal{G}$ as exactly the same way of constructing $\mathcal{P}^{\text{trunc}}$ for $\mathcal{P}$. Therefore, by Lemma H.7, we have

$$\begin{aligned}&\mathbb{E}_\pi^{\mathcal{P}}[\|\mathbb{P}^{\mathcal{G}}(\cdot,\cdot\,|\,c_h)-\mathbb{P}^{\mathcal{G}^{\text{trunc}}}(\cdot,\cdot\,|\,c_h)\|_1]\\&\le\mathbb{E}_\pi^{\mathcal{P}}\sum_{t\in[h]}\|\mathbb{T}_t(\cdot\,|\,s_t,a_t)-\mathbb{T}_t^{\text{trunc}}(\cdot\,|\,s_t,a_t)\|_1+\|\mathbb{O}_t(\cdot\,|\,s_t)-\mathbb{O}_t^{\text{trunc}}(\cdot\,|\,s_t)\|_1\\&\le4\mathbb{P}^{\pi,\mathcal{G}}(\exists t\in[h]:s_t\notin\mathcal{S}_t^{\text{high}})\\&\le4HS\epsilon_1.\end{aligned}$$

Meanwhile, we can construct $\widehat{\mathcal{G}}^{\text{trunc}}$ and $\widehat{\mathcal{G}}^{\text{sub}}$ using the exactly way as for $\widehat{\mathcal{P}}^{\text{trunc}}$ and $\widehat{\mathcal{P}}^{\text{sub}}$, where $\widehat{\mathcal{G}}^{\text{sub}}$ is an γ-observable POSGs.

Now, according to [43], for all examples in Appendix C.4, there exists a compression function that maps c_h to $\widehat{c}_h$ such that the size of the compressed common information is quasi-polynomial, i.e., $\widehat{C}_h \leq (AO)^{C\gamma^{-4}\log\frac{SH}{\epsilon_2}}$ for some absolute constant C, and corresponding approximate belief $\{\widehat{P}_h : \widehat{\mathcal{C}}_h \to \Delta(\mathcal{S}\times\mathcal{P}_h)\}_{h\in[H]}$ such that

$$\mathbb{E}_{\pi}^{\widehat{\mathcal{G}}^{\text{sub}}} \|\mathbb{P}^{\widehat{\mathcal{G}}^{\text{sub}}}(\cdot,\cdot \mid c_h) - \widetilde{P}_h(\cdot,\cdot \mid \widehat{c}_h)\|_1 \leq \epsilon_2.$$

Therefore, we can do the same augmentation for $\widetilde{P}_h$ on states from $\mathcal{S}_h^{\text{low}}$ to construct the approximate belief $\widehat{P}_h$ as in the proof of Theorem H.5, and the remaining steps are exactly the same as the proof of Theorem H.5. This will lead to a total of polynomial-time and polynomial-sample complexities. ■

J.1 Background on Bayesian Games

The Bayesian game is a generalization of normal-form games in partially observable settings. Specifically, a bayesian game is specified as $(n, \{\mathcal{A}_i\}_{i\in[n]}, \{\Theta_i\}_{i\in[n]}, \{r_i\}_{i\in[n]}, \mu)$, where n is the number of players, $\mathcal{A}_i$ is the actor space, Θ_i is the type space, $r_i : \Theta\times\mathcal{A}\to[0,1]$ is the reward function, and μ is the prior distribution of the joint type. At the beginning of the game, a type $\theta = (\theta_i)_{i\in[n]}$ is drawn from the prior distribution $\mu\in\Delta(\Theta)$. Then each agent i gets its only type θ_i and take the action a_i. With a slight abuse of the notation, we define a strategy of an agent as $\gamma_i\in\Gamma_i = \{\Theta_i\to\Delta(\mathcal{A}_i)\}$. We define $J_i(\gamma_i,\gamma_{-i})$ to be the expected rewards for agent i, given the joint strategy γ_i,γ_{-i}.

By definition, $J_i(\gamma_i,\gamma_{-i})$ can be evaluated as

$$J_i(\gamma_i,\gamma_{-i}) := \mathbb{E}_{\theta\sim\mu}\mathbb{E}_{\{a_j\sim\gamma_j(\cdot\mid\theta_j)\}_{j\in[n]}} r_i(\theta,a).$$

Bayesian NE. We define $\gamma^\star$ is an ϵ-NE is it satisfies that

$$J_i(\gamma^\star) \geq J_i(\gamma_i', \gamma_{-i}^\star) - \epsilon, \forall i\in[n], \gamma_i'\in\Gamma_i.$$

Bayesian CCE. We say a distribution of joint strategies $s\in\Delta(\Gamma)$ to be a ϵ-Bayesian CCE if it satisfies

$$\mathbb{E}_{\gamma\sim s} J_i(\gamma) \geq \mathbb{E}_{\gamma\sim s} J_i(\gamma_i',\gamma_{-i}) - \epsilon, \forall i\in[n], \gamma_i'\in\Gamma_i.$$

(Agent-form) Bayesian CE. We say a distribution of joint strategies $s\in\Delta(\Gamma)$ to be an ϵ-agent-form Bayesian CE if it satisfies

$$\mathbb{E}_{\gamma\sim s} J_i(\gamma) \geq \mathbb{E}_{\gamma\sim s} J_i(m_i\diamond\gamma_i,\gamma_{-i}) - \epsilon, \forall i\in[n], m_i'\in\mathcal{M}_i,$$

where $\mathcal{M}_i = \{\Theta_i\times\mathcal{A}_i\to\mathcal{A}_i\}$ is the space for strategy modification, where m_i modifies γ_i as follows: given current type θ_i and the recommended action a_i, the strategy modification changes the action to the another action $m_i(\theta_i,a_i)$.

Note that Bayesian NE for zero-sum games, and (agent-form) Bayesian CE/CCE are all tractable solution concepts and can be computed with polynomial computation complexity, e.g., [25, 28, 19].

K Concluding Remarks and Limitations

In this paper, we aim to understand the provable benefits of privileged information for partially observable RL problems under two empirically successful paradigms, with an emphasis on *both computational and sample efficiency* of the algorithms. We summarized our results in Table 1, showing that privileged information does yield a significant improvement for a series of well-known POMDP subclasses. One potential limitation of our work is that we only focused on the case with *exact* state information. It remains to explore whether such an assumption can be further relaxed, e.g., when privileged state information is biased, partially observable, or delayed, as usually happens in practice, and how our theoretical results may be affected. Meanwhile, as an initial theoretical study, we have been primarily focusing on the tabular settings, and it would be interesting to extend the results to more function-approximate settings.