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, xyliu999@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 has 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. 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 asymmetric actor-critic algorithm with polynomial sample and quasi-polynomial computational complexities, in which one key component is a new provable oracle for learning belief states that preserves *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 featuring *centralized-training-with-decentralized-execution*, a popular framework in empirical MARL, with polynomial sample and (quasi-)polynomial computational complexities 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 [46, 3], autonomous driving [73, 41], dialogue systems [88], and clinical trials [77], only *partial observations* of the environment state are available for sequential decision-making. Such partial observability presents significant challenges for efficient decision-making and learning, with known computational [66] and statistical [43, 36] 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 not fully-observable in decision-making [85, 80].

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 known as

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

 https://doi.org/10.52202/079017-2038

the *centralized-training-with-decentralized-execution* (CTDE) framework in deep MARL, and has become prevalent in practice [53, 70, 22, 82]. 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 some empirical successes in robotic locomotion [45, 59] and autonomous driving [14]; 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* algorithm [68], and serves as the backbone of some high-profiled successes in robotic manipulation [46, 3] and MARL [53, 82].

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* [44, 30] (see Appendix B for a detailed discussion). However, most of these theoretically sound algorithms are different from those used in practice, and require 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 [44, 30], 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 was recently shown to allow provable partially observable RL without computationally intractable oracles [25]. 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 summary). Further, we revisit the *asymmetric actor-critic* paradigm and analyze its efficiency under the more challenging setting of observable POMDPs above (where expert distillation fails). Identifying the inefficiency of vanilla asymmetric actor-critic, and inspired by the empirical success in *belief-state-learning*, we develop a new *belief-weighted* 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 multi-agent RL 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. Consider a POMDP characterized 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 transition matrices, so that $\mathbb{T}_h(\cdot \,|\, s, a) \in \Delta(\mathcal{S})$ 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 reaching each next state from different current states. 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 $\mathbb{O} = \{\mathbb{O}_h\}_{h\in[H]}$ to denote the collection of emission matrices, so that $\mathbb{O}_h(\cdot \,|\, s) \in \Delta(\mathcal{O})$ gives the emission distribution over the 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 $\mathbb{O}_h(\cdot \,|\, s)$ for each $s \in \mathcal{S}$. Finally, $r = \{r_h\}_{h\in[H]}$ is a collection of reward functions, so that $r_h(s, a) \in [0, 1]$ is the reward given the state s and action a at step h. When privileged information is available, the agent can observe the underlying state $s \in \mathcal{S}$ directly *during training* (only). We thus denote the trajectory until step h *with states* as $\overline{\tau}_h = (s_{1:h}, o_{1:h}, a_{1:h-1})$, the one *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 underlying state at step h given history τ_h, which is known as the *belief state* (c.f. Appendix C.1 for more details).

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

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

where the agent bases on the entire (partially observable) history for decision-making. The corresponding policy class is denoted as Π_h. We further denote $\Pi = \times_{h\in[H]}\Pi_h$. We also 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: fix an integer $L > 0$, we define the policy space Π^L to be the space of all possible policies $\pi = \pi_{1:H} := (\pi_h)_{h\in[H]}$ 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 state-based policies as $\Pi_{\mathcal{S}}$, i.e., for any $\pi = \pi_{1:H} \in \Pi_{\mathcal{S}}$, $\pi_h : \mathcal{S} \to \Delta(\mathcal{A})$ for all $h \in [H]$.

Given the POMDP model $\mathcal{P}$, we write $\mathbb{P}^{\mathcal{P}}_{s_{1:H+1},a_{1:H},o_{1:H}\sim\pi}(\mathcal{E})$ to denote the event $\mathcal{E}$ when $(s_{1:H+1}, a_{1:H}, o_{1:H})$ is drawn as a trajectory following the policy π in the model $\mathcal{P}$. We will also use the shorthand notation $\mathbb{P}^{\pi,\mathcal{P}}(\cdot)$ if $(s_{1:H+1}, a_{1:H}, o_{1:H})$ is evident. We write $\mathbb{E}^{\mathcal{P}}_{\pi}[\cdot]$ to denote the expectation similarly. We define the value function at step h as $V^{\pi,\mathcal{P}}_h(y_h) := \mathbb{E}^{\mathcal{P}}_{\pi}[\sum_{t=h}^{H} r_t(s_t, a_t) \,|\, y_h]$, denoting the expected accumulated rewards from step h, where $y_h \subseteq (s_{1:h}, o_{1:h}, a_{1:h-1})$, and we slightly abuse the 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 to the agent 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 $v^{\mathcal{P}}(\pi) = \mathbb{E}^{\mathcal{P}}_{\pi}[\sum_{h=1}^{H} r_h(s_h, a_h)]$. Meanwhile, we also define $Q^{\pi,\mathcal{P}}_h(y_h, a_h) := \mathbb{E}^{\mathcal{P}}_{\pi}[\sum_{t=h}^{H} r_t(s_t, a_t) \,|\, y_h, a_h]$. We denote the occupancy measure on the state space as $d^{\pi,\mathcal{P}}_h(s_h) = \mathbb{P}^{\pi,\mathcal{P}}(s_h)$. The goal of learning in POMDPs is to find the optimal policy that *maximizes* the expected accumulated reward over the policies that take τ_h as input at each step $h \in [H]$, i.e., those $\pi \in \Pi$. 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* the training and test time, the agent can only observe its historical observations and actions τ_h at step h, 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 training. 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, and it is the performance for such policies that matters eventually. For simplicity, we assume the reward function is known since under our privileged information setting, learning the reward function is much easier than learning the transition and emission, and the sample/computational complexity for the former is dominated by that for the latter. This assumption has also been made for learning in POMDPs without privileged information [36, 47, 48].

2.2 Partially Observable Multi-agent RL 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 $r_i = \{r_{i,h}\}_{h\in[H]}$ with $r_{i,h}(s, a) \in [0, 1]$ denoting the reward given state s and joint action a for agent i at step h. An episode of POSG proceeds as follows: at each step h and state s_h, a joint 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 the corresponding action $a_{i,h}$, obtains the reward $r_{i,h}(s_h, a_h)$, where $a_h := (a_{i,h})_{i\in[n]}$, and then the system transitions to the next state as $s_{h+1} \sim \mathbb{T}_h(\cdot \,|\, s_h, a_h)$. 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. The joint private information at step h is denoted as $p_h = (p_{i,h})_{i\in[n]}$, where the collection of the joint private information is given by $\mathcal{P}_h = \mathcal{P}_{1,h} \times \cdots \times \mathcal{P}_{n,h}$. We refer more examples of this setting of POSG with information-sharing to Appendix C.2 (and also [62, 63, 51]). 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

https://doi.org/10.52202/079017-2038

	Without PI	With PI (Ours)
Block MDP	With STD: Oracle-efficient [17, 18, 60] Without additional assump.: Computationally harder than SL [28]	Tabular: Poly sample + time
k-decodable POMDP	Exponential-in-k sample + time [19]	FA: Poly sample + Classification (SL) oracle
Det. POMDP	Without WSE: Statistically hard [47] With WSE: Poly sample + time [36]	
POSG with det. filter	N/A	Poly sample + time
Observable POMDP	Quasi-poly sample + time [25] [51]	Poly sample + Quasi-poly time
Observable POSG		

Table 1: Comparison of the theoretical guarantees with and without privileged information. PI: privileged information; STD: structural assumptions on transition dynamics, e.g., deterministic transition or reachability of all states; SL: supervised learning; FA: function approximation; WSE: well-separated emission.

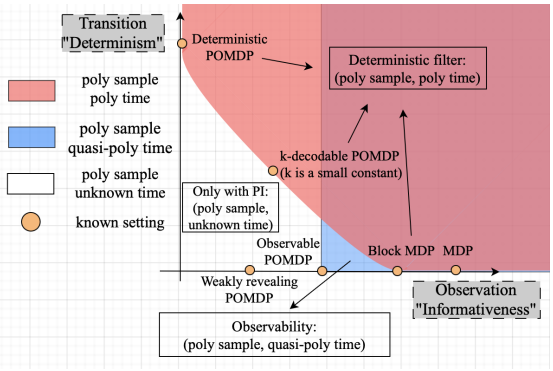

Figure 1: A landscape of POMDP models that partially observable RL with privileged information can/cannot address. The axes denote the "restrictiveness" of the assumptions, on the emission channels and transition dynamics, respectively.

space for agent i as Π_i. If $\pi_{i,h}$ takes the state s_h instead of $(c_h, p_{i,h})$ as input, we denote its policy space as $\Pi_{\mathcal{S},i}$, e.g., for each agent i, and policy $\pi_{1:H} \in \Pi_{\mathcal{S},i}$, we have $\pi_{i,h} : \mathcal{S} \to \Delta(\mathcal{A}_i)$ for each step $h \in [H]$. Similar to the POMDP setting, we define Π^{gen} to be the most general policy space, i.e., $\Pi^{\text{gen}} := \{\pi_{1:H} \mid \pi_h : \mathcal{S}^h \times \mathcal{O}^h \times \mathcal{A}^{h-1} \to \Delta(\mathcal{A}) \text{ for } h \in [H]\}$. Note that this model covers several recent POSG models studied for partially observable MARL, e.g., [49, 27]. For example, at each step h, if there is no shared information, then $c_h = \emptyset$, and if all history information is shared, then $p_{i,h} = \emptyset$ for all $i \in [n]$. 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. The solution concepts for POSGs are usually the *equilibria*, particularly Nash equilibrium (NE) for two-player zero-sum games (i.e., when $n = 2$ and $r_{1,h} + r_{2,h} = 1$),[1] and correlated equilibrium (CE) or coarse correlated equilibrium (CCE) for general-sum games. We defer the formal definitions of these standard solution concepts to Appendix C.2.

2.3 Technical Assumptions for Computational Tractability

A key technical assumption is that the POMDPs/POSGs we consider satisfy an *observability* assumption, as 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 [26, 25, 51]. Meanwhile, we defer an additional assumption to ensure the traceability for solving POSGs to Appendix C.3.

Assumption 2.2 (γ-observability [20, 26, 25])**.** 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.

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 the privileged information as input. This is usually referred to as either *expert distillation* [14, 64, 58] or *teacher-student learning* [45, 59, 75] 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 partial observation (history) as input. One prominent example of ii) is *asymmetric-actor-critic* [68, 3]. 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* [53, 86, 21], which is usually instantiated under the actor-critic framework, with the critic taking privileged information as input

[1] Note that we require $r_{1,h} + r_{2,h}$ to be 1 instead of 0 to be consistent with our assumption that $r_{i,h} \in [0, 1]$ for each $i \in [0, 1]$, and this requirement does not lose optimality as one can always subtract the constant-sum offset to attain a zero-sum reward structure.

in training. Here we formalize and revisit the potential pitfalls of these two paradigms, and further develop theoretical guarantees under certain additional conditions and/or algorithm variants.

3.1 Privileged Policy Learning: Expert Policy Distillation

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

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

where $\pi' \in \Pi^{\text{gen}}$ is some given behavior 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 algorithm. However, we point out in the following proposition that even if the POMDP satisfies Assumption 2.2, the distilled policy can still be strictly suboptimal even with infinite data, i.e., by solving the expected objective Equation (3.1) completely. 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 a γ-observable POMDP $\mathcal{P}^\epsilon$ with $H = 1$, $S = O = A = 2$ such that for any behavior policy $\pi' \in \Pi^{\text{gen}}$ 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 Equation (3.1) fails is that in general, the underlying state can remain highly uncertain even given the full history. Thus, the distilled policy may not be able to mimic the state-based expert policy well at different states s_h if the associated $\pi^\star_h(\cdot \mid s_h)$ differs significantly across s_h. To see how we may rule out such an issue, notice that if $\gamma = 1$ (note that according to Assumption 2.2, we have γ is at most 1 since $\gamma \le \|\mathbb{O}_h\|_\infty \le 1$), implying that the observation can decode the underlying state, the bound in Proposition 3.1 becomes vacuous. Inspired by this, we propose the following condition that incorporates 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 $h \ge 2$, the belief update operator under $\mathcal{P}$ satisfies that there exists an *unknown 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 $b^s(s) = 1$ is 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, $B_1(\mu_1; o_1) = b^{\psi_1(o_1)}$, where $B_h(b; o_h) := \mathbb{P}^{\mathcal{P}}_{s_h\sim b}(\cdot \mid o_h) \in \Delta(\mathcal{S})$, $U_h(b; a_{h-1}, o_h) := \mathbb{P}^{\mathcal{P}}_{s_{h-1}\sim b}(\cdot \mid a_{h-1}, o_{h-1}) \in \Delta(\mathcal{S})$ are the belief update operators under the Bayes rule for any $b \in \Delta(\mathcal{S})$, 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 including Block MDP, deterministic POMDP, k-decodable POMDP as well as a new setting we have identified and existing literature cannot handle. We refer the formal introduction to Appendix E and Figure 1 for an illustration.

In light of the pitfall in Proposition 3.1, we will analyze *both* the computational and statistical efficiencies of expert distillation in Section 4, under the condition in Definition 3.2.

3.2 Privileged Value Learning: Asymmetric Actor-Critic

Asymmetric actor-critic [68] iterates between two procedures as in standard actor-critic algorithms [42]: policy *improvement* and policy *evaluation*. As the name suggests, its key difference from the standard actor-critic is that the algorithm maintains Q-value functions (the critic) based on the *state/privileged information*, while the policy receives only the (partially observable) *history* as input.

Policy evaluation. At iteration $t-1$, given the policy π^{t-1}, the algorithm estimates Q-functions in the form of $\{Q_h^{t-1}(\tau_h, s_h, a_h)\}_{h\in[H]}$, where we adopt the "unbiased" version [6] such that Q-

 https://doi.org/10.52202/079017-2038

functions are conditioned on *both* the *history* and the *states*. [2] One key to achieving sample efficiency is by adding some bonus terms in policy evaluation to encourage exploration, i.e., obtaining some *optimistic* Q-function estimates, similarly as in the fully-observable MDP setting, see e.g., [11, 74], for which we defer the detailed introduction to Section 4.

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

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

where λ_t is the step-size and PROJ_Π is the projection operator onto the space of Π, which corresponds to projecting onto the simplex of $\Delta(\mathcal{A})$ for each $h \in [H]$. Here we point out the potential drawback of the vanilla algorithm as in [68, 6], where the key insight is that for each iteration of policy evaluation and improvement, one roughly *only* performs the computation of order $\mathcal{O}(KH)$, while needing to collect K new episodes of samples. Thus, the *sample complexity* will scale in the same order as the *computational complexity* when the algorithm converges after some iterations to an ϵ-optimal solution, which will be super-polynomial even for γ-observable POMDPs [27]. Proof of the result is deferred to Appendix E.

Proposition 3.3 (Inefficiency of vanilla asymmetric actor-critic)**.** Under the tabular parameterization for both the policy and the value function, 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 an issue, one may need to perform *more* computation *per iteration*, so that although the *total* computational complexity (iteration number $\times$ per-iteration computational complexity) is super-polynomial, the total iteration number can be lower such that the total sample complexity may be lower as well. This desideratum is hard to achieve if one *computes* policy update only on the *sampled* trajectories τ_h per iteration, i.e., update *asynchronously*, since this will couple the scales of computational and sample complexities similarly as Equation (3.2). In contrast, we first propose to update *all* trajectories per iteration in a *synchronous* way, with the following proximal-policy optimization-type [72] policy improvement update with the state-history-dependent Q-functions $\{Q_h^{t-1}(\tau_h, s_h, a_h)\}_{h\in[H]}$:

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

where we recall $\boldsymbol{b}_h(\tau_h) \in \Delta(\mathcal{S})$ denotes the belief state and $\eta > 0$ is the learning rate. This update rule also reduces to the natural policy gradient (NPG) [40] update under the softmax policy parameterization in the fully-observable case [1], when updated for each state s_h separately [74]. We defer the detailed derivation of Equation (3.3) to Appendix E.

However, such an update presents two challenges: (1) It requires enumerating all possible τ_h, whose number scales exponentially with the horizon, making it still computationally intractable; (2) An explicit belief function $\boldsymbol{b}_h$ is needed. Motivated by these two challenges, 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 efficiently (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 approximate belief learning. Such a decoupling enables a more modular algorithm design framework, and can potentially incorporate the rich literature on learning approximate beliefs in practice, see e.g., [24, 65, 16, 87, 83], which has mostly not been theoretically analyzed before. We will thus analyze such an oracle in Section 5.

[2] As pointed out in [6], the original asymmetric actor-critic [68], where the value function was only conditioned on the *state*, is a *biased* estimate of the actual history-conditioned value function that appears in the policy gradient in the infinite-horizon discounted setting. We verify that such a state-based value function is indeed also biased under our finite-horizon setting, see Remark E.1 for an example. We will thus use the unbiased value function estimate conditioned on *both* the state and the history throughout.

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) underlying 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 $\mathcal{M}$ derived from POMDP $\mathcal{P}$ by omitting the observations and observing the underlying state (see Definition 4.2 for a formal definition), and for each step $h \in [H]$, a decoding function $g_h(s_{h-1}, a_{h-1}, o_h)$ such that the probability of *incorrectly* decoding a state-action-observation triplet over the trajectories induced by the policy π^E is low.

Definition 4.2 (POMDP-induced MDP)**.** Given a POMDP $\mathcal{P} = (H, \mathcal{S}, \mathcal{A}, \mathcal{O}, \mathbb{T}, \mathbb{O}, \mu_1, r)$, we define its associated Markov Decision Process (MDP) $\mathcal{M}$ as $\mathcal{M} = (H, \mathcal{S}, \mathcal{A}, \mathbb{T}, \mu_1, r)$ without observations.

Definition 4.3. Consider a POMDP $\mathcal{P}$ that satisfies Definition 3.2, and let $\psi = \{\psi_h\}_{h\in[H]}$ be the promised set of functions that always correctly decode a state-action-observation triplet into an underlying state. Consider policy $\widetilde{\pi^E} = \{\pi_h^E(\psi(\cdot)) : \mathcal{S} \times \mathcal{A} \times \mathcal{O} \to \Delta(\mathcal{A})\}_{h\in[H]} \in \Pi^D$. We slightly abuse the notation and simply denote by $v^{\mathcal{P}}(\pi^E) = v^{\mathcal{P}}(\widetilde{\pi^E})$.

Lemma 4.4. 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 the 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 an approximate optimal policy π^E for the associated MDP $\mathcal{M}$ (see Definition 4.2). Thus, in the rest of the section, we focus on learning the decoding function $\{g_h\}_{h\in[H]}$. To efficiently learn the decoding function, we model the access to the underlying state by keeping track of the most recent pair of the action and the observation, as well as the two most recent states. We summarize the algorithm of decoding-function learning in Algorithm 1.

Theorem 4.5. 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 output of Algorithm 1 with $M = \frac{AOS+\log(H/\delta)}{\epsilon^2}$. Then, with probability at least $1-\delta$, for each step $h \in [H]$: $\mathbb{P}^{\pi^E,\mathcal{P}}\left[\exists h \in [H] : g_h\left(s_{h-1}, a_{h-1}, o_h\right) \neq s_h\right] \leq \epsilon$, using $\text{POLY}(H, A, O, S, \frac{1}{\epsilon}, \log\left(\frac{1}{\delta}\right))$ 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 4.4 and Theorem 4.5. Note that *both* the sample and computation complexities are *polynomial*, which is in stark contrast to the k-decodable POMDP case [19] (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 [19]. 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 using privileged information in several other subclasses of problems, we refer to Table 1 for more details.

Theorem 4.6. Let $\mathcal{P}$ satisfy Definition 3.2 and consider any policy $\pi^E \in \Pi_{\mathcal{S}}$. Using $\text{POLY}(H, A, O, S, \frac{1}{\epsilon}, \log\left(\frac{1}{\delta}\right))$ episodes and in time $\text{POLY}(H, A, O, S, \frac{1}{\epsilon}, \log\left(\frac{1}{\delta}\right))$, we can compute a policy $\pi \in \Pi^D$ (see Definition 4.1) such that with probability at least $1-\delta$, $v^{\mathcal{P}}(\pi) \geq v^{\mathcal{P}}(\pi^E) - \epsilon$.

Extension to the case with general function approximation. Due to the modularity of our algorithmic framework and its compatibility 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.

https://doi.org/10.52202/079017-2038

5 Provable Asymmetric Actor-Critic with Approximate Belief Learning

Unlike most existing theoretical studies on provably sample-efficient partially observable RL [36, 25, 47], 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 a *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 methods [40, 1, 74] in the fully-observable setting, with additional weighting over the states s_h using some learned belief states, to handle the additional *state*-dependence in the asymmetric critic. The overall algorithm is presented in Algorithm 2. The algorithm requires a belief-learning subroutine that takes the stored memory as input and outputs a belief about the underlying 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 establish the performance guarantee of Algorithm 2 in the following theorem. We defer the proof to Appendix H.

Theorem 5.1 (Near-optimal policy)**.** Fix $\epsilon, \delta \in (0,1)$. 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 at least $1-\delta$, Algorithm 2 can learn an approximate 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 $\text{POLY}(S, A, O, H, Z, \frac{1}{\epsilon}, \log\frac{1}{\delta})$, where ϵ_{belief} is the belief-learning error defined as $\epsilon_{\text{belief}} := \max_{h\in[H]} \max_{\pi\in\Pi^L} \mathbb{E}_\pi^{\mathcal{P}} \|\boldsymbol{b}_h(\tau_h) - \boldsymbol{b}_h^{\text{apx}}(z_h)\|_1$ and $Z := \max_{h\in[H]} |\mathcal{Z}_h|$. Furthermore, if $\mathcal{P}$ is additionally γ-observable (c.f. Assumption 2.2), then $\pi^\star$ is also an approximate 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 Approximate Belief Learning

At a high level, our belief-learning algorithm first learns an approximate POMDP model $\widehat{\mathcal{P}}$ by explicitly exploring the state space. The main technical challenge here is that there may exist states that are reachable with very low probability, making it infeasible to collect enough samples to sufficiently explore them, thus potentially breaking the γ-observability property of the ground-truth model $\mathcal{P}$. To circumvent this issue, we ignore such hard-to-visit states and *redirect* probabilities flowing to them to certain other states. Thus, in our truncated POMDP, where each state is sufficiently explored, we can approximate the transition and emission matrices to a desired accuracy *uniformly across all the states* and preserve the γ-observability property. This ensures that the learned approximate belief function in the truncated POMDP is sufficiently close to the actual belief function of the original POMDP $\mathcal{P}$. Note that the key to achieving belief learning with *both* polynomial sample and time complexities is our explicit exploration in 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 [44]. However, for such exploration to be *computationally tractable*, one requires to avoid using computationally intractable oracles for *POMDP learning*, which is in fact our final goal. We present the guarantees in the next theorem and postpone the proof to Appendix H.

Theorem 5.2. Consider a γ-observable POMDP $\mathcal{P}$ (c.f. Assumption 2.2) and assume that $L \geq \widetilde{\Omega}(\gamma^{-4}\log(SH/\epsilon))$ for an $\epsilon > 0$. Then, we can learn an approximate belief $\{\boldsymbol{b}_h^{\text{apx}}\}_{h\in[H]}$ from Algorithm 4 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 at least $1-\delta$, 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 guarantee below. In contrast to the case without privileged information [25, 27], the sample complexity is reduced from quasi-polynomial to polynomial for γ-observable POMDPs. Note that the computational complexity remains quasi-polynomial, which is known to be unimprovable even for planning [27]. The key to such an improvement, as pointed out in Section 3.2, 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* finite-memory). 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.

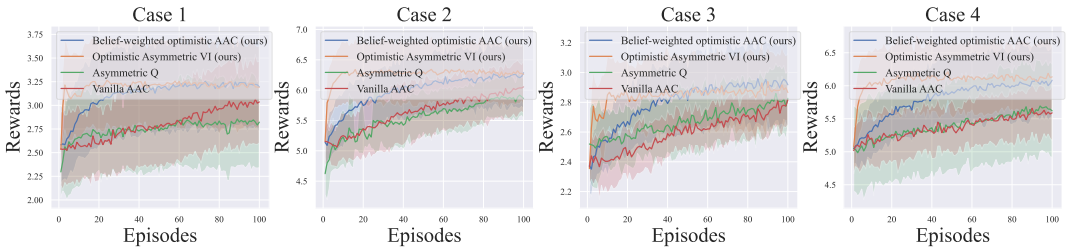

Figure 2: Results for POMDPs of different sizes, where our methods achieve the best performance with the lowest sample complexity (VI: value iteration; AAC: asymmetric actor-critic).

Theorem 5.3. Let $\mathcal{P}$ be a γ-observable POMDP (c.f. Assumption 2.2), and consider $L \geq \widetilde{\Omega}(\gamma^{-4}\log(SH/\epsilon))$ for an $\epsilon > 0$. With probability at least $1-\delta$, Algorithm 2 can learn a policy $\pi \in \Pi^L$ such that $v^{\mathcal{P}}(\pi) \geq \max_{\pi' \in \Pi} v^{\mathcal{P}}(\pi') - \epsilon$, using $\mathrm{POLY}(S, H, 1/\epsilon, 1/\gamma, \log(1/\delta), O, A)$ episodes and in time $\mathrm{POLY}(S, H, 1/\epsilon, \log(1/\delta), O^L, A^L)$.

6 Numerical Validation

We now provide some numerical results for both of our principled algorithms. Here we mainly compare with two baselines, the vanilla asymmetric actor-critic [68], and asymmetric Q-learning [7], on two settings, POMDP under the deterministic filter condition (c.f. Definition 3.2) and general POMDPs. We report the results in Table 2 and Figure 2, where our algorithms converge faster to higher rewards. We defer the implementation details and discussions to Appendix I.

7 Extension to Partially Observable MARL with Privileged Information

7.1 Privileged Policy Learning: Equilibrium Distillation

To understand how the deterministic filter condition may be extended for POSGs, we first note the following equivalent characterization of Definition 3.2, the proof of which is deferred to Appendix J.

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

Proposition 7.1 implies that at each step h, given the *entire* history information, the agent can uniquely decode the current underlying state s_h. Thus, we generalize this condition to POSGs by requiring each agent to uniquely decode the current state s_h given the information it has collected 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 *an unknown* function $\phi_{i,h} : \mathcal{C}_h \times \mathcal{P}_{i,h} \to \mathcal{S}$ 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 underlying state through their own information *individually*. Naturally, one may wonder whether one can relax it so that only the *joint* history information of all the agents can decode the underlying state. However, we point out in the following that it does not circumvent the computational hardness of POSG, the proof of which is deferred to Appendix J. 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, with which one can simulate the RL problems with privileged information.

Proposition 7.3. Computing CCE in POSGs that satisfy that for each step $h \in [H]$, there exists a function $\phi_h : \mathcal{C}_h \times \mathcal{P}_h \to \mathcal{S}$ 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, our framework for POSGs is also decoupled into two steps: i) learning an *expert* equilibrium policy that is fully observable, ii) learning the *decoding function*, where the first step can be instantiated by any provable off-the-shelf algorithm of learning in 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 distilled student policy is bounded by the 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 each agent i to make sure the decoding function is accurate under policies

https://doi.org/10.52202/079017-2038

(π_i', π_{-i}) for any π_i', keeping π_{-i} fixed. We refer the detailed algorithm to Algorithm 5, and present below the guarantees for learning the decoding functions and the corresponding distilled policy for learning NE/CCE, while we defer the results for learning CE to Theorem J.6.

Theorem 7.4 (Equilibria learning; Combining Theorem J.1 and Theorem J.4)**.** Under Assumption C.8 and conditions of Definition 7.2, given a $\frac{\epsilon}{2}$-NE/CCE π^E for the associated Markov game of $\mathcal{G}$, Algorithm 5 can learn decoding function $\{\widehat{g}_{i,h}\}_{i\in[n],h\in[H]}$ such that with probability at least $1-\delta$, 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 complexities $\mathrm{POLY}(S, A, H, O, \frac{1}{\epsilon}, \log\frac{1}{\delta})$. Consequently, policy π distilled from π^E (c.f. Theorem J.1 for the formal distillation procedures) 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 computational efficiency. We generalize to POSGs with information sharing by defining the *compression* of the common information.

Definition 7.5 (Compressed approximate common information [57, 79, 51])**.** For each $h\in[H]$, given a set $\widehat{\mathcal{C}}_h$, we say $\mathrm{Compress}_h$ is 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 condition 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 $c_{h+1} := c_h \cup \varpi_{h+1}$ and the definition of Υ_{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 [63, 62, 51], 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. Note that both $\widehat{P}_h$ and thus ϵ_{belief} have implicit dependencies on $\mathrm{Compress}_h$, as $\widehat{c}_h := \mathrm{Compress}_h(c_h)$. We outline our algorithm in Algorithm 7, which is conceptually similar to the algorithm for POMDPs (Algorithm 2), maintaining the asymmetric critic (i.e., value function), and performing the actor update (i.e., 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.8, with probability at least $1-\delta$, 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 computational 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 computational complexity is still to firstly learn approximate models, i.e., transitions and emissions, and then carefully *truncate* (as in Section 5.2) its transition and emission to build the approximate belief, where we defer the detailed algorithm to Algorithm 8. Next, we provide its provable guarantees, which leads to a final polynomial-sample and quasi-polynomial-time complexity result when combined with Theorem 7.6.

Theorem 7.7. For any $\epsilon > 0$, under Assumption 2.2, 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 computational complexity $\mathrm{POLY}(S, A, O, H, \frac{1}{\gamma}, \frac{1}{\epsilon}, \log\frac{1}{\delta})$ for all the examples in Appendix C.3. As a consequence, Algorithm 7 can learn an ϵ-NE if $\mathcal{G}$ is zero-sum and ϵ-CE/CCE if $\mathcal{G}$ is general-sum with sample complexity $\mathcal{O}(\frac{H^4SAO\log(SAHO/\delta)}{\epsilon^2})$ and computational complexity $\mathrm{POLY}(S, (AO)^{\mathcal{O}(\gamma^{-4}\log(SH/\epsilon))}, H, \frac{1}{\epsilon}, \log\frac{1}{\delta})$.

Acknowledgement

The authors would like to thank the anonymous reviewers and area chair from NeurIPS 2024 for the valuable feedback. Y.C. acknowledges the support from the NSF Awards CCF-1942583 (CAREER) and CCF-2342642. X.L. and K.Z. acknowledge the support from Army Research Laboratory (ARL) Grant W911NF-24-1-0085. K.Z. also acknowledges the support from Simons-Berkeley Research Fellowship. A.O. acknowledges financial support from a Meta PhD fellowship, a Sloan Foundation Research Fellowship and the NSF Award CCF-1942583 (CAREER). Part of this work was done while the authors were visiting the Simons Institute for the Theory of Computing.

References

[1] A. Agarwal, S. M. Kakade, J. D. Lee, and G. Mahajan. Optimality and approximation with policy gradient methods in Markov decision processes. In *Conference on Learning Theory*, pages 64–66, 2020.

[2] 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.

[3] 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.

[4] R. J. Aumann, M. Maschler, and R. E. Stearns. *Repeated games with incomplete information*. MIT press, 1995.

[5] 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.

[6] 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.

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

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

[9] 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.

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

[11] Q. Cai, Z. Yang, C. Jin, and Z. Wang. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pages 1283–1294. PMLR, 2020.

[12] 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.

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

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

[15] 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.

 https://doi.org/10.52202/079017-2038

[16] 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.

[17] C. Dann, N. Jiang, A. Krishnamurthy, A. Agarwal, J. Langford, and R. E. Schapire. On oracle-efficient pac rl with rich observations. *Advances in neural information processing systems*, 31, 2018.

[18] S. Du, A. Krishnamurthy, N. Jiang, A. Agarwal, M. Dudik, and J. Langford. Provably efficient rl with rich observations via latent state decoding. In *International Conference on Machine Learning*, pages 1665–1674. PMLR, 2019.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

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

[24] 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.

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

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

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

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

[33] 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.

[34] N. Jiang. On value functions and the agent-environment boundary. *arXiv preprint arXiv:1905.13341*, 2019.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] S. Kakade and J. Langford. Approximately optimal approximate reinforcement learning. In *International Conference on Machine Learning*, volume 2, pages 267–274, 2002.

[40] S. M. Kakade. A natural policy gradient. In *Advances in Neural Information Processing Systems*, pages 1531–1538, 2002.

[41] B. R. Kiran, I. Sobh, V. Talpaert, P. Mannion, A. A. Al Sallab, S. Yogamani, and P. Pérez. Deep reinforcement learning for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 23(6):4909–4926, 2021.

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

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

[44] 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.

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

[46] 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.

[47] 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.

[48] 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.

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

[50] 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.

[51] 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.

 https://doi.org/10.52202/079017-2038

[52] X. Liu and K. Zhang. Partially observable multi-agent reinforcement learning with information sharing, 2024.

[53] 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.

[54] 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.

[55] 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.

[56] 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.

[57] 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.

[58] 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.

[59] 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.

[60] 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.

[61] 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.

[62] 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.

[63] 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.

[64] 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.

[65] 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.

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

[67] 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.

[68] L. Pinto, M. Andrychowicz, P. Welinder, W. Zaremba, and P. Abbeel. Asymmetric actor critic for image-based robot learning. *Robotics: Science and Systems XIV*, 2018.

[69] 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.

[70] 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.

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

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

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

[74] 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.

[75] 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.

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

[77] 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.

[78] 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.

[79] 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.

[80] 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.

[81] 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.

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

[83] 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.

[84] 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.

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

[86] 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.

 https://doi.org/10.52202/079017-2038

[87] 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.

[88] 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.

[89] A. Zanette and E. Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*, pages 7304–7312. PMLR, 2019.

[90] 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

1 Introduction **1**

2 Preliminaries **2**

2.1 Partially Observable RL (with Privileged Information) 2

2.2 Partially Observable Multi-agent RL with Information Sharing 3

2.3 Technical Assumptions for Computational Tractability 4

3 Revisiting Empirical Paradigms of RL with Privileged Information **4**

3.1 Privileged Policy Learning: Expert Policy Distillation 5

3.2 Privileged Value Learning: Asymmetric Actor-Critic 5

4 Provably Efficient Expert Policy Distillation **7**

5 Provable Asymmetric Actor-Critic with Approximate Belief Learning **8**

5.1 Belief-Weighted Optimistic Asymmetric Actor-Critic 8

5.2 Approximate Belief Learning 8

6 Numerical Validation **9**

7 Extension to Partially Observable MARL with Privileged Information **9**

7.1 Privileged Policy Learning: Equilibrium Distillation 9

7.2 Privileged Value Learning: Asymmetric MARL with Approximate Belief Learning 10

A Societal Impact **19**

B Related Work **19**

C Additional Preliminaries **20**

C.1 Additional Preliminaries on POMDPs 20

C.2 Additional Preliminaries for POSGs 20

C.2.1 Evolution of the Common and Private Information 23

C.3 Strategy Independence of Belief and Examples 23

D Collection of Algorithms **24**

E Missing Details in Section 3 **31**

F Missing Details in Section 4 **33**

https://doi.org/10.52202/079017-2038

G Provably Efficient Expert Policy Distillation with Function Approximation **34**

H Missing Details in Section 5 **36**

H.1 Supporting Technical Lemmas 44

I Missing Details in Section 6 **47**

J Missing Details in Section 7 **48**

J.1 Background on Bayesian Games 61

K Concluding Remarks and Limitations **62**

A Societal Impact

Our work is theoretical by nature, 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 [43] and computationally intractable [66], a productive line of research has identified several structured subclasses of POMDPs that can be efficiently solved. [43] introduced the class of POMDPs in the rich-observation setting, where the observation space can be large and fully reveal the underlying state, where sample-efficient RL becomes possible [35, 60]. [19] introduced k-step decodable POMDPs, where the last k observation-action pairs can uniquely determine the state, and proposed polynomial-sample complexity algorithms (assuming k is a small constant). Beyond settings where the underlying state can be *exactly* recovered, [36, 47] 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., [84, 15, 12, 54, 48, 90]. Notably, these algorithms are typically computationally inefficient, requiring access to an optimistic planning oracle for POMDPs. On a promising note, [26] showed that in observable POMDPs (see Assumption 2.2), one can have quasi-polynomial-time algorithms for planning the near-optimal policy, which further leads to provable RL algorithms [25, 27] with *both quasi-polynomial* samples and time.

RL under hindsight observability. The closest line of research to ours are the recent theoretical studies for Hindsight Observable Markov Decision Processes (HOMDPs) [44], where the underlying state is revealed at the end of the episode; see also subsequent related works in [30, 76] 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 settings. 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` . [66]. Meanwhile, even under the additional assumption of observability on the *underlying* POMDP model, it is still not clear if these algorithms can avoid computationally intractable oracles, since the approximate POMDP that [44] needs to do planning on at every iteration during learning can be quite different from the underlying model. For example, at the beginning of exploration when not 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 underlying POMDP. This makes that single iteration even computationally intractable. In contrast, our focus is on better understanding practically inspired algorithmic paradigms, without computationally intractable oracles, which in practice often do have and use the privileged state information *during* each episode (instead of only at the end) [68, 45, 14].

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 [14, 64, 58], also known as *teacher-student* learning [45, 59, 75], as we analyze in Section 4. For the latter, asymmetric actor-critic [68] represents one of the well-known examples, with other studies in [7, 3]. 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., [53, 22, 70, 86, 82]. Intriguingly, it was shown that if the privileged value function depends *only on* the state, the associated actor will cause bias [6, 55, 56]. This has thus necessitated the use of history/belief in asymmetric actor-critic, as in our Section 5. Notably, the empirical framework in [83] 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

https://doi.org/10.52202/079017-2038

belief-state learning and planning [24, 65, 33, 16, 87] as we study in Section 5, oftentimes with privileged state information to supervise the belief learning procedure [61, 5].

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 a *belief* over the underlying state by the historical observations and actions.

Definition C.1 (Belief state update)**.** For each $h \in [H+1]$, the Bayes operator $B_h : \Delta(\mathcal{S}) \times \mathcal{O} \to \Delta(\mathcal{S})$ 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 $x \in \mathcal{S}$. For each $h \in [H+1]$, 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 $(a_{1:h}, o_{1:h})$, 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 [26] and define

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

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 $(a_{h-L:h}, o_{h-L+1:h})$, we inductively define

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

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

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 agent i 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 $a \in \mathcal{A}$ is taken at state $s \in \mathcal{S}$ 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 from different current states. We use μ_1 to denote the distribution of the initial state s_1, and $\mathcal{O}_i$ to denote the observation space for agent i 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 $\mathbb{O} = \{\mathbb{O}_h\}_{h \in [H]}$ 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

$\mathbb{O}_{i,h}(\cdot \mid s) \in \Delta(\mathcal{O}_i)$ as the marginalized emission for agent i 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 agent i 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 such history at step h is denoted by $\mathcal{T}_{i,h}$.

In many practical examples, agents may share part of the history with each other, which may introduce some *information structures* of the game that may lead to both sample and computation efficiencies. 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 $\tau_h = (\tau_{i,h})_{i\in[n]}$: $c_h \subseteq \{o_1, a_1, o_2, \cdots, a_{h-1}, o_h\}$, 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 her private information $p_{i,h} = \tau_{i,h} \setminus c_h$, where the collection of the private information for agent i is denoted as $\mathcal{P}_{i,h}$ and its cardinality as $P_{i,h}$. The cardinality of the joint private information 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, where the goal is usually to find the team-optimal solution.

Policies and value functions. We define a stochastic policy for agent i at step h as:

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

where Ω_h is the random seed space, which is shared among agents and $\omega_{i,h} \in \Omega_h$ is the random seed for agent i. 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 agent i, i.e., the private information together with the common information, to the distribution over her actions.

We define π_i as a sequence of policies for agent i at all steps $h \in [H]$, i.e., $\pi_i = (\pi_{i,1}, \cdots, \pi_{i,H})$. We further denote $\Pi_i = \times_{h\in[H]}\Pi_{i,h}$ as the policy space for agent i and $\Pi = \prod_{i\in[n]} \Pi_i$ as the joint policy space. As a special case, we define the space of *deterministic* policy as $\widetilde{\Pi}_i$, where $\widetilde{\pi}_i \in \widetilde{\Pi}_i$ maps the private information and common information to a *deterministic* action for agent i and the joint space as $\widetilde{\Pi} = \prod_{i\in[n]} \widetilde{\Pi}_i$.

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, and a (potentially correlated) joint policy is denoted as $\pi = \pi_1 \odot \pi_2 \cdots \odot \pi_n \in \Pi$.

We are now ready to define the *value function* conditioned on the common information under our model of POSG with information sharing:

Definition C.2 (Value function with information sharing)**.** For each agent $i \in [n]$ and step $h \in [H]$, given common information c_h and joint policy $\pi = (\pi_i)_{i=1}^n \in \Pi$, the *value function conditioned on the common information* of agent i is defined as: $V_{i,h}^{\pi,\mathcal{G}}(c_h) := \mathbb{E}_\pi^{\mathcal{G}}\left[\sum_{h'=h}^H r_{i,h'}(s_{h'}, a_{h'}) \,\middle|\, c_h\right]$, 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 $h-1$,

 https://doi.org/10.52202/079017-2038

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 all agents' private information p_h. With the common-information-based conditional belief, the value function given in Definition C.2 has the following recursive structure:

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

where the expectation is taken over the randomness of (s_h, p_h, a_h, o_{h+1}). With this relationship, we can define the *prescription-value* function correspondingly, a generalization of the *action-value* function in (fully observable) stochastic games and MDPs to POSGs with information sharing, as follows.

Definition C.3 (Prescription-value function with information sharing)**.** 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 agent i is defined as:

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

where prescription $\gamma_{i,h} \in \Delta(\mathcal{A}_i)^{P_{i,h}}$ replaces the partial function $\pi_{i,h}(\cdot \mid \omega_{i,h}, c_h, \cdot)$ in the value function. From now on, we will refer to it as *prescription-value function* for short. With such a prescription-value function, agents can take actions purely based on their local private information [63, 62, 51].

This prescription-value function indicates the expected return for agent i 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. For a joint policy $(\pi_i)_{i=1}^n \in \Pi$ we denote the expected reward of agent i by $v_i^{\mathcal{G}}(\pi) = \mathbb{E}_{\pi}^{\mathcal{P}}[\sum_{h=1}^H r_{i,h}(s_h, a_h)]$.

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* for agent i, 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 $m_i \diamond \pi_i$ is defined as follows: at step h, when agent i is given c_h and $p_{i,h}$, the joint 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 [78, 50, 38, 49, 51] when there exists certain common information, and is a natural generalization of the definition in the normal-form game case [71]. We denote by $\mathcal{M}_i^{\text{gen}}$ the space of all possible strategy modifications m_i if it conditions on any history information instead of only $(c_h, p_{i,h})$. Similarly, we use $\mathcal{M}_{\mathcal{S},i}$ to denote the space of all possible strategy modifications m_i if it only conditions on the current state e.g., a modification $m_{i,h,s} : \mathcal{A}_i \to \mathcal{A}_i$.

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 over 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 obtained 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 [62, 63], specifying the source of common information and private information. Based on such evolution rules, we define $\{f_h\}_{h\in[H]}$ and $\{g_h\}_{h\in[H]}$, 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 $h \in [H]$, as the mappings that map the joint history to common information and joint private information, respectively.

C.3 Strategy Independence of Belief and Examples

To solve a POSG without computationally intractable oracles, certain information-sharing is needed even under the observability assumption [51]. We thus make the following assumption as in [51].

Assumption C.8 (Strategy independence of beliefs [62, 31, 51])**.** 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 $p_h \in \mathcal{P}_h, s_h \in \mathcal{S}$, $\mathbb{P}_h^{\pi_{1:h-1},\mathcal{G}}(s_h, p_h \,|\, c_h) = \mathbb{P}_h^{\pi'_{1:h-1},\mathcal{G}}(s_h, p_h \,|\, c_h)$.

We provide examples satisfying this assumption in Appendix C.2, which include the fully-sharing structure as in [27, 69] as a special case. Finally, we also assume that common information and private information evolve over time properly in Assumption C.7, as standard in [62, 63, 51], which covers the models considered in [27, 69, 49].

Here we take the examples from [52, 51] to illustrate the generality of the information-sharing framework.

Example C.9 (One-step delayed sharing)**.** At any step $h \in [H]$, 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 [2].

Example C.10 (State controlled by one controller with asymmetric delay sharing)**.** We assume there are 2 players for convenience. It extends naturally to n-player settings. Consider the case where the state dynamics are controlled by player 1, i.e., $\mathbb{T}_h(\cdot \,|\, s_h, a_{1,h}, a_{2,h}) = \mathbb{T}_h(\cdot \,|\, s_h, a_{1,h}, a'_{2,h})$ for all $(s_h, a_{1,h}, a_{2,h}, a'_{2,h}, h)$. There are two kinds of delay-sharing structures we could consider: **Case A:** the information structure is given as $c_h = \{o_{1,2:h}, o_{2,2:h-d}, a_{1,1:h-1}\}$, $p_{1,h} = \emptyset$, $p_{2,h} = \{o_{2,h-d+1:h}\}$, i.e., player 1's observations are available to player 2 instantly, while player 2's observations are available to player 1 with a delay of $d \geq 1$ time steps. **Case B:** similar to **Case A** but player 1's observation is available to player 2 with a delay of 1 step. The information structure is given as $c_h = \{o_{1,2:h-1}, o_{2,2:h-d}, a_{1,1:h-1}\}$, $p_{1,h} = \{o_{1,h}\}$, $p_{2,h} = \{o_{2,h-d+1:h}\}$, where $d \geq 1$. This kind of asymmetric sharing is common in network routing [67], where packages arrive at different hosts with different delays, leading to asymmetric delay sharing among hosts.

https://doi.org/10.52202/079017-2038

Example C.11 (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.12 (Information sharing with one-directional-one-step delay)**.** Similar to the previous cases, we also assume there are 2 players for ease of exposition, and the case can be generalized to multi-player cases straightforwardly. Similar to the one-step delay case, we consider the situation where all observations of player 1 are available to player 2, while the observations of player 2 are available to player 1 with one-step delay. All the past actions are available to both players. That is, in this case, $c_h = \{o_{1,2:h}, o_{2,2:h-1}, a_{1:h-1}\}$, 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.13 (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 [4].

D Collection of Algorithms

Algorithm 1 Learning Decoding Function with Privileged Information

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

Ensure: A decoding function for each step $\{g_h\}_{h\in[H]}$ (see Theorem 4.6)

For the $h = 1$ step: Collect M state-observation pairs from the first-step $\widehat{D}_1 = \left\{\left(s_1^{(i)}, o_1^{(i)}\right)\right\}_{i\in[M]}$ on POMDP $\mathcal{P}$ and define the decoding function g_1 for the first step as:

$$g_1(o_1) = \left\{s_1 : (s_1, o_1) \in \widehat{D}_1\right\}$$

for each step $h \in [2, H]$ **do**

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

$$\widehat{D}_h := \left\{\left(s_{h-1}^{(i)}, a_{h-1}^{(i)}, o_h^{(i)}, s_h^{(i)}\right)\right\}_{i\in[M]}$$

Define the decoding function g_h for step 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 $\{g_h\}_{h\in[H]}$

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

Require:
- POMDP $\mathcal{P}$,
- Subroutine T_Q that given policy $\pi \in \Pi^L$, outputs $\{\widetilde{Q}\}_{h\in[H]}$ 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**

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

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

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

Denote $\pi^t = \pi_{1:H}^t$

end for

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

Algorithm 3 Optimistic Q-function Estimation with Privileged Information

Require:
- POMDP $\mathcal{P}$, policy $\pi_{1:H} \in \Pi^L$ such that $\pi_h : \mathcal{S} \to \Delta(\mathcal{A})$,
- Number of episodes M per step.

Ensure: Approximate Q-functions $\{\widetilde{Q}_h\}_{h\in[H]}$ (see Lemma H.3)

Initialize: $\forall z_{H+1} \in \mathcal{Z}_{H+1}, s_{H+1} \in \mathcal{S}, a_{H+1} \in \mathcal{A}$

$$\widetilde{Q}_{H+1}(z_{H+1}, s_{H+1}, a_{H+1}) \leftarrow 0, \qquad \pi_{H+1}(a_{H+1} \mid z_{H+1}) \leftarrow \frac{1}{A}$$

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

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

Compute empirical counts and define empirical distributions:

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

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

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

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

where $\widetilde{V}_{h+1}(z_{h+1}, s_{h+1}) = \mathbb{E}_{a_{h+1}\sim\pi_{h+1}(\cdot \mid z_{h+1})}[\widetilde{Q}_{h+1}(z_{h+1}, s_{h+1}, a_{h+1})]$

end for

end for

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

https://doi.org/10.52202/079017-2038

Algorithm 4 Approximate Belief Learning with Privileged Information via Model Truncation

Require:

- POMDP $\mathcal{P} = (H, \mathcal{S}, \mathcal{A}, \mathcal{O}, \{\mathbb{T}_h\}_{h\in[H]}, \{\mathbb{O}_h\}_{h\in[H]}, \mu_1, \{r_h\}_{h\in[H]})$,
- An MDP learning oracle MDP_Learning that efficiently learns an approximate optimal policy of an MDP,
- Number of trajectories N,
- The threshold ϵ.

Ensure: Approximate belief $\{\boldsymbol{b}_h^{\text{apx}}\}_{h\in[H]}$ (See Theorem H.5)

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 the first $h-1$ steps then take action a_h for each $a_h \in \mathcal{A}$ deterministically and denote the dataset $\{(s_h^i, o_h^i, a_h^i, s_{h+1}^i)\}_{i\in[NA]}$

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

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

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

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

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

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

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

end for

end for

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

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

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

end for

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

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

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

$\widehat{\mu}_1^{\text{trunc}}(s_1) := \widehat{\mu}_1(s_1) + \frac{\sum_{s'_1\in\mathcal{S}_1^{\text{low}}} \widehat{\mu}_1(s'_1)}{|\mathcal{S}_1^{\text{high}}|}, \forall s_1 \in \mathcal{S}_1^{\text{high}}$

end for

Let

$$\widehat{\mathcal{P}}^{\text{sub}} := (H, \{\mathcal{S}_h^{\text{high}}\}_{h\in[H]}, \mathcal{A}, \mathcal{O}, \{\widehat{\mathbb{T}}_h^{\text{trunc}}\}_{h\in[H]}, \{\widehat{\mathbb{O}}_h^{\text{trunc}}\}_{h\in[H]}, \widehat{\mu}_1^{\text{trunc}}, \{r_h\}_{h\in[H]})$$

Define $\{\widehat{\boldsymbol{b}}_h^{\prime,\text{sub}} : \mathcal{Z}_h \to \Delta(\mathcal{S}_h^{\text{high}})\}_{h\in[H]}$ to be the approximate belief w.r.t. $\widehat{\mathcal{P}}^{\text{sub}}$

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

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

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

Require:

- POSG $\mathcal{G} = (H, \mathcal{S}, \mathcal{A}, \mathcal{O}, \{\mathbb{T}_h\}_{h\in[H]}, \{\mathbb{O}_h\}_{h\in[H]}, \mu_1, \{r_i\}_{i\in[n]})$,
- $\pi \in \Pi_{\mathcal{S}}$, controller set $\{\mathcal{I}_h \subseteq [n]\}_{h\in[H]}$,
- Procedure MDP_Learning$(\cdot,\cdot)$ that takes as input an MDP and a reward function and returns an approximate optimal policy,
- Number of trajectories N.

Ensure: A decoding function for each agent and step $\{\widehat{g}_{j,h}\}_{j\in[n],h\in[H]}$ (see Theorem J.4)

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 the Markov game $\mathcal{M}$ associated with $\mathcal{G}$ as $\mathcal{M} = (H, S, A, \{\mathbb{T}_h\}_{h\in[H]}, \mu_1)$, where we omit the specification for the reward functions and one can specify them arbitrarily
    Define $\mathcal{M}(\pi_{-i})$ 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
  For each $i \in [n]$, $a_h \in \mathcal{A}$, collect N trajectories by executing policy $\Psi_i(h, s_h) \times \pi_{-i}$ for the first $h-1$ steps then take action a_h deterministically 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 $(o_h, a_h, s_{h+1}) \in \mathcal{O} \times \mathcal{A} \times \mathcal{S}$ **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 $\{\widehat{g}_{j,h}\}_{j\in[n],h\in[H]}$

https://doi.org/10.52202/079017-2038

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

Require:

- POSG $\mathcal{G} = (H, \mathcal{S}, \mathcal{A}, \mathcal{O}, \{\mathbb{T}_h\}_{h\in[H]}, \{\mathbb{O}_h\}_{h\in[H]}, \mu_1, \{r_i\}_{i\in[n]})$,
- $\pi \in \Pi_{\mathcal{S}}$, 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,
- Number of trajectories N.

Ensure: A decoding function for each agent and step $\{\widehat{g}_{j,h}\}_{j\in[n],h\in[H]}$ (see Theorem J.6)

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 $\mathcal{M}_i^{\text{extended}}(\pi)$ to be the *extended* MDP, which is defined in Definition J.5.
    $\Psi_i(h, s_h) \leftarrow$ MDP_Learning$(\mathcal{M}_i^{\text{extended}}(\pi), \widehat{r}_i)$
  end for
  For each $i \in [n]$, $a_h \in \mathcal{A}$, collect N trajectories by executing policy $\Psi_i(h, s_h) \times \pi_{-i}$ for the first $h-1$ steps then take action a_h deterministically 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 $(o_h, a_h, s_{h+1}) \in \mathcal{O} \times \mathcal{A} \times \mathcal{S}$ **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 $\{\widehat{g}_{j,h}\}_{j\in[n],h\in[H]}$

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

Require:
- POSG $\mathcal{G} = (H, \mathcal{S}, \mathcal{A}, \mathcal{O}, \{\mathbb{T}_h\}_{h\in[H]}, \{\mathbb{O}_h\}_{h\in[H]}, \mu_1, \{r_i\}_{i\in[n]})$,
- An approximate belief $\{\widehat{P}_h : \widehat{\mathcal{C}}_h \to \Delta(\mathcal{S}\times\mathcal{P}_h)\}_{h\in[H]}$,
- Number of iterations K.

Ensure: An approximate equilibrium policy

Initialize:

$$N_h^0(s_h, a_h) \leftarrow 0, \quad N_h^0(s_h, a_h, o_h) \leftarrow 0, \quad \widehat{\mathbb{J}}^0(o_{h+1} \mid s_h, a_h) \leftarrow \frac{1}{O}$$

for $k \in [K]$ **do**
  for $h \leftarrow H, H-1, \cdots, 1$ **do**
    for $\widehat{c}_h \in \widehat{\mathcal{C}}_h$ **do**

$$Q_{i,h}^{\text{high},k}(\widehat{c}_h, p_h, s_h, a_h) \leftarrow \min\left\{r_{i,h}(s_h,a_h) + b_h^{k-1}(s_h,a_h) + \mathbb{E}_{o_{h+1}\sim\widehat{\mathbb{J}}_h^{k-1}(\cdot\mid s_h,a_h)}\left[V_{i,h+1}^{\text{high},k}(\widehat{c}_{h+1})\right], H-h+1\right\} \text{ for } i\in[n]$$

$$Q_{i,h}^{\text{low},k}(\widehat{c}_h, p_h, s_h, a_h) \leftarrow \max\left\{r_{i,h}(s_h,a_h) - b_h^{k-1}(s_h,a_h) + \mathbb{E}_{o_{h+1}\sim\widehat{\mathbb{J}}_h^{k-1}(\cdot\mid s_h,a_h)}\left[V_{i,h+1}^{\text{low},k}(\widehat{c}_{h+1})\right], 0\right\} \text{ for } i\in[n]$$

Define

$$Q_{i,h}^{\text{high},k}(\widehat{c}_h, \gamma_h) := \mathbb{E}_{s_h,p_h\sim\widehat{P}_h(\cdot,\cdot\mid\widehat{c}_h)}\mathbb{E}_{\{a_{j,h}\sim\gamma_{j,h}(\cdot\mid p_{j,h})\}_{j\in[n]}}\left[Q_{i,h}^{\text{high},k}(\widehat{c}_h, p_h, s_h, a_h)\right] \text{ for } i\in[n]$$

Define

$$Q_{i,h}^{\text{low},k}(\widehat{c}_h, \gamma_h) := \mathbb{E}_{s_h,p_h\sim\widehat{P}_h(\cdot,\cdot\mid\widehat{c}_h)}\mathbb{E}_{\{a_{j,h}\sim\gamma_{j,h}(\cdot\mid p_{j,h})\}_{j\in[n]}}\left[Q_{i,h}^{\text{low},k}(\widehat{c}_h, p_h, s_h, a_h)\right] \text{ for } i\in[n]$$

$\{\pi_{j,h}^k(\cdot\mid\cdot,\widehat{c}_h,\cdot)\}_{j\in[n]} \leftarrow$ Bayesian-CE/CCE$(\{Q_{j,h}^{\text{high},k}(\widehat{c}_h,\cdot)\}_{j\in[n]})$ (c.f. Appendix J.1)

$V_{i,h}^{\text{high},k}(\widehat{c}_h) \leftarrow \mathbb{E}_{\omega_h}\left[Q_{i,h}^{\text{high},k}(\widehat{c}_h, \{\pi_{j,h}^k(\cdot\mid\omega_{j,h},\widehat{c}_h,\cdot)\}_{j\in[n]})\right]$ for $i\in[n]$

$V_{i,h}^{\text{low},k}(\widehat{c}_h) \leftarrow \mathbb{E}_{\omega_h}\left[Q_{i,h}^{\text{low},k}(\widehat{c}_h, \{\pi_{j,h}^k(\cdot\mid\omega_{j,h},\widehat{c}_h,\cdot)\}_{j\in[n]})\right]$ for $i\in[n]$

    end for
  end for
  Execute π^k and get trajectory $(s_{1:H}^k, a_{1:H}^k, o_{1:H+1}^k)$
  for $h\in[H], s_h\in\mathcal{S}, a_h\in\mathcal{A}, o_{h+1}\in\mathcal{O}$ **do**
    $N_h^k(s_h,a_h) \leftarrow \sum_{l\in[k]}\mathbb{1}[s_h^l = s_h, a_h^l = a_h]$
    $N_h^k(s_h,a_h,o_{h+1}) \leftarrow \sum_{l\in[k]}\mathbb{1}[s_h^l = s_h, a_h^l = a_h, o_{h+1}^l = o_{h+1}]$
    $\widehat{\mathbb{J}}_h^k(o_{h+1}\mid s_h,a_h) \leftarrow \frac{N_h^k(s_h,a_h,o_{h+1})}{N_h^k(s_h,a_h)}$
  end for
end for
$k^\star \leftarrow \arg\min_{k\in[K], i\in[n]} V_{i,1}^{\text{high},k}(c_1^k) - V_{i,1}^{\text{low},k}(c_1^k)$
return $\pi^{k^\star}$ for general-sum games or the marginalized policy of $\pi^{k^\star}$ for zero-sum games

https://doi.org/10.52202/079017-2038

Algorithm 8 Approximate Belief Learning for MARL with Privileged Information

Require:
- POSG $\mathcal{G} = (H, \mathcal{S}, \mathcal{A}, \mathcal{O}, \{\mathbb{T}_h\}_{h\in[H]}, \{\mathbb{O}_h\}_{h\in[H]}, \mu_1, \{r_{i,h}\}_{i\in[n],h\in[H]})$,
- Controller set $\{\mathcal{I}_h \subseteq [n]\}_{h\in[H]}$,
- Procedure MDP_Learning$(\cdot,\cdot)$,
- Number of trajectories N,
- Accuracy ϵ.

Ensure: An approximate belief

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 $(h', s'_h, a'_h) \in [H] \times \mathcal{S} \times \mathcal{A}$.

$\mathcal{M} \leftarrow (H, \mathcal{S}, \mathcal{A}, \{\mathbb{T}_h\}_{h\in[H]}, \mu_1)$ to be the MDP by ignoring the observation and emission of $\mathcal{G}$, where we omit the specification for the reward functions and one can specify them arbitrarily

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

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

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

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

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

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

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

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

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

end for

end for

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

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

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

end for

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

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

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

$\widehat{\mu}_1^{\text{trunc}}(s_1) := \widehat{\mu}_1(s_1) + \frac{\sum_{s'_1\in\mathcal{S}_1^{\text{low}}} \widehat{\mu}_1(s'_1)}{|\mathcal{S}_1^{\text{high}}|}, \forall s_1 \in \mathcal{S}_1^{\text{high}}$

end for

Let

$$\widehat{\mathcal{G}}^{\text{sub}} := (H, \{\mathcal{S}_h^{\text{high}}\}_{h\in[H]}, \mathcal{A}, \mathcal{O}, \{\widehat{\mathbb{T}}_h^{\text{trunc}}\}_{h\in[H]}, \{\widehat{\mathbb{O}}_h^{\text{trunc}}\}_{h\in[H]}, \widehat{\mu}_1^{\text{trunc}}, \{r_{i,h}\}_{i\in[n],h\in[H]})$$

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

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

return $\{\widehat{P}_h\}_{h\in[H]}$

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 underlying state, see Appendix C for a detailed introduction. Note that Equation (3.1) can be written as

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

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:

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

Now we are ready to construct the counter-example of γ-observable POMDP $\mathcal{P}^\epsilon$ for some $\epsilon\in(0,1)$ with $H=1$, $\mathcal{S}=\{s^1,s^2\}$, $\mathcal{A}=\{a^1,a^2\}$, and $\mathcal{O}=\{o^1,o^2\}$. We let $\mu_1=(\frac{1-\gamma}{2-\gamma},\frac{1}{2-\gamma})$, $\mathbb{O}_1(o^1\,|\,s^1)=1$, and $\mathbb{O}_1(o^1\,|\,s^2)=1-\gamma$, $\mathbb{O}_1(o^2\,|\,s^2)=\gamma$. 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 $\pi_1^\star(a^1\,|\,s^1)=1$ and $\pi_1^\star(a^2\,|\,s^2)=1$, i.e., choosing a^1 at state s^1 and a^2 at state s^2 deterministically. Meanwhile, by our construction, one can compute that the belief given observation o^1 ensures $\boldsymbol{b}_1(o^1)=\mathrm{Unif}(\mathcal{S})$. Hence, the corresponding "distilled" partially observable policy under observation o^1 is given by

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

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 $\widehat{\pi}_1^\star(\cdot\,|\,o_1)=\mathrm{Unif}(\mathcal{A})$. On the other hand, combining the fact that $\boldsymbol{b}_1(o^1)=\mathrm{Unif}(\mathcal{S})$ with $\epsilon<1$, it is direct to see that the optimal partially observable policy $\widetilde{\pi}\in\arg\max_{\pi\in\Pi} v^{\mathcal{P}}(\pi)$ satisfies $\widetilde{\pi}_1(a^1\,|\,o^1)=1$. Now we are ready to evaluate the optimality gap between $\widetilde{\pi}$ and $\widehat{\pi}^\star$ as follows

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

where the last step is due to the fact that $\widetilde{\pi}$ is the optimal policy, leading to the fact that $V_1^{\widetilde{\pi},\mathcal{P}^\epsilon}(o^2)-V_1^{\widehat{\pi}^\star,\mathcal{P}^\epsilon}(o^2)\geq 0$. 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. ■

https://doi.org/10.52202/079017-2038

Note that another counterexample in a similar spirit was also constructed in [34], demonstrating that the expert policy for a poorly chosen agent-environment boundary can be useless in imitation learning, although the γ-observability property is not satisfied for the construction therein.

Remark E.1. The counter-example $\mathcal{P}^\epsilon$ constructed above can be also used to demonstrate the *bias* of the state-only-based value function as an estimate of the history-based value function that appeared in the policy gradient formula for POMDPs, in the finite-horizon setting, i.e. $\mathbb{E}_{s_h \sim \boldsymbol{b}_h(\tau_h)}[V_h^{\pi,\mathcal{P}^\epsilon}(s_h)] \neq V_h^{\pi,\mathcal{P}^\epsilon}(\tau_h)$ (mirroring Theorem 4.2 of [6]). Specifically, in the counter-example above, we consider the policy π such that $\pi_1(a_1 \,|\, o_1) = 1$ and $\pi_1(a_2 \,|\, o_2) = 1$. The state-only-based value function can be evaluated as

$$V_1^{\pi,\mathcal{P}^\epsilon}(s_1) = \frac{1-\gamma}{2-\gamma}, \qquad V_1^{\pi,\mathcal{P}^\epsilon}(s_2) = \gamma\epsilon,$$

which implies that $\mathbb{E}_{s_1 \sim \boldsymbol{b}_1(o_1)}[V_1^{\pi,\mathcal{P}^\epsilon}(s_1)] = \frac{V_1^{\pi,\mathcal{P}^\epsilon}(s_1)+V_1^{\pi,\mathcal{P}^\epsilon}(s_2)}{2} = \frac{\frac{1-\gamma}{2-\gamma}+\gamma\epsilon}{2}$. On the other hand, it holds that $V_1^{\pi,\mathcal{P}^\epsilon}(o_1) = \frac{1+\epsilon}{2}$, which is not equal to $\mathbb{E}_{s_1 \sim \boldsymbol{b}_1(o_1)}[V_1^{\pi,\mathcal{P}^\epsilon}(s_1)]$, showing the bias of such a state-only value function.

Example E.2 (Deterministic POMDP [36, 81])**.** 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.3 (Block MDP [43, 18])**.** 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 $\mathrm{supp}(\mathbb{O}_h(\cdot \,|\, s_h)) \cap \mathrm{supp}(\mathbb{O}_h(\cdot \,|\, s_h')) = \emptyset$ when $s_h \neq s_h'$.

Example E.4 (k-step decodable POMDP [19])**.** We say a POMDP $\mathcal{P}$ is a 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, $\mathbb{P}^{\mathcal{P}}(s_h = \phi_h^\star(z_h) \,|\, \tau_h) = 1$, where $\mathcal{Z}_h = (\mathcal{O}\times\mathcal{A})^{\max\{h-1,k-1\}} \times \mathcal{O}$, $z_h = ((o,a)_{k(h):h-1}, o_h)$, and $k(h) = \max\{h-k+1, 1\}$.

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.4 to be unknown and arbitrary (instead of being a small known constant as in [19] to have provably efficient algorithms).

Example E.5 (POMDP with arbitrary, unknown decodable length)**.** This example is similar to Example E.4, but the decoding length m is unknown and not necessarily a small constant.

In light of the pitfall in Proposition 3.1, we will analyze *both* the computational and statistical efficiencies of expert distillation in Section 4, under the condition in Definition 3.2.

Proof of Example E.2 & Example E.3 & Example E.4 & Example E.5: 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 computational complexity and sample complexity of each iteration t as follows.

Sample complexity: The algorithm executes the policy π^{t-1} and collect K episodes, denoted as $\{s_{1:H+1}^k, o_{1:H}^k, a_{1:H}^k\}_{k\in[K]}$. Thus, the sample complexity of each iteration is $\Theta(K)$.

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

$$Q_h^t(\tau_h^k, s_h^k, a_h^k) \leftarrow (1-\alpha) Q_h^{t-1}(\tau_h^k, s_h^k, a_h^k) + \alpha \left(r_h(s_h^k, a_h^k) + \frac{1}{|\mathcal{J}(\tau_h^k, s_h^k, a_h^k)|} \sum_{j \in \mathcal{J}(\tau_h^k, s_h^k, a_h^k)} Q_{h+1}^t(\tau_{h+1}^j, s_{h+1}^j, a_{h+1}^j) \right),$$

for some stepsize $\alpha \in (0,1)$, where $\mathcal{J}(\tau_h^k, s_h^k, a_h^k) := \{j \in [K] \,|\, (\tau_h^j, s_h^j, a_h^j) = (\tau_h^k, s_h^k, a_h^k)\}$. Therefore, the computational complexity for this procedure is of $\mathrm{POLY}(H, K)$.

Computational complexity for policy improvement: For tabular parameterization, computing $\nabla \log \pi_h^{t-1}(a_h^k \,|\, \tau_h^k)$ 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 [26]. This implies that the vanilla asymmetric actor-critic needs to take super-polynomial time 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 the Q-function not depend on the entire history τ_h but only the finite-memory z_h, the proof still holds. The key is that whenever one only *computes* at the *sampled* history/finite-memories, i.e., updates the policy in an *asynchronous* way (in contrast to the *synchronous* one where the policies at *all* histories/finite-memories are updated), the sample and computational complexities will be coupled with the same order per iteration, which implies a super-polynomial sample complexity due to the super-polynomial computational complexity. This completes the proof. ■

Derivation for the closed-form update Equation (3.3). Note that the proximal policy optimization [72] update has the policy improvement as follows

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

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}:

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

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

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

F Missing Details in Section 4

Proof of Lemma 4.4: The proof follows by the assumption that the total cumulative reward is at most H,

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

which completes the proof. ■

https://doi.org/10.52202/079017-2038

Proof of Theorem 4.5: For each step $h \in [H]$ we define D_h to be the distribution over the underlying state s_{h-1} at step $h-1$, taken action $a_{h-1} \in \mathcal{A}$ based on π^E, the underlying state transitioned to $s_h \in \mathcal{S}$, and the observation $o_h \sim \mathbb{O}_h(\cdot \mid s_h)$. We remind that we include at step zero, a dummy state-observation pair (s_0, o_0), for notational convenience. Formally, D_h is defined as $D_h(s_{h-1}, a_{h-1}, o_h, s_h) := \mathbb{P}^{\pi^E, \mathcal{P}}[s_{h-1}, a_{h-1}, o_h, s_h]$.

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

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

For each $h \in [H]$, we can use M episodes to collect M samples from the distribution D_h. In addition, since state s_{H+1} is dummy, we need not to collect episodes for D_{H+1}. Denote the set of collected samples by $\widehat{D}_h^M$. We define the decoding g_h for step $h \in [H]$ as follows:

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

Observe that by Definition 3.2, $\{s_h \mid (s_{h-1}, a_{h-1}, o_h, s_h) \in \widehat{D}_h^M\}$ is either the empty set or contains only a single elements, in which case, it is true that $g_h(s_{h-1}, a_{h-1}, o_h) = \psi_h(s_{h-1}, a_{h-1}, o_h)$ (ψ is the real decoding function, see Definition 3.2). Moreover, we slightly abuse the notation and let $\widetilde{D}_h^M$ denote the empirical distribution induced by the samples in $\widehat{D}_h^M$. Thus, with probability at least $1 - \frac{\delta}{H}$ and setting $M = \Theta\left(\frac{A \cdot O \cdot S + \log(H/\delta)}{\epsilon^2}\right)$ for each step $h \in [H]$, we obtain the following by the result in [13]:

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

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

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

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

G Provably Efficient Expert Policy Distillation with Function Approximation

We now turn our attention to the rich-observation setting under our 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 step $h \in [H]$, 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 O, for each step $h \in [H]$, 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 $\mathcal{F} = \{\mathcal{F}_h\}_{h\in[H]}$ 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 the PAC learnability for multi-class classification [8]. Defining the DS dimension is beyond the scope of our paper; we direct interested readers to [8] 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 [8])**.** 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 $(x, y) \in \text{supp}(D)$, $g^*(x) = y$. Given n samples from D, there exists an algorithm that with probability at least $1 - \delta$ outputs $\widetilde{g} \in \mathcal{F}$ such that

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

where $d_{DS}(\mathcal{F})$ denotes the Daniely and Shalev-Shwartz-Dimension of the function class $\mathcal{F}$.

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 [9], there exists an algorithm learning the decoding function $\{g_h\}_{h\in[H]}$ such that with probability at least $1-\delta$, for each step $h \in [H]$:

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

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 Daniely and Shalev-Shwartz-Dimension of $\mathcal{F}_h$ [8].

Combining Theorem G.3 and Lemma 4.4, we obtain the final polynomial sample complexity in this function approximation setting, using classification (supervised learning) oracles (c.f. Table 1).

Proof of Theorem G.3:

For each step $h \in [H]$, we define D_h to be the distribution over the underlying state s_{h-1} at step $h-1$, taken action $a_{h-1} \in \mathcal{A}$ from π^E, the underlying state transitioned to $s_h \in \mathcal{S}$, and the hallucinated observation $o_h \sim \mathbb{O}_h(\cdot \mid s_h)$ (we remind readers that for step 0, we use dummy state s_0 and action a_0). Formally, the probability that the sequence $(s_{h-1}, a_{h-1}, o_h, s_h)$ is sampled from D_h equals to

$$D_h(s_{h-1}, a_{h-1}, o_h, s_h) := \mathbb{P}^{\pi^E,\mathcal{P}}\left[s_{h-1}, a_{h-1}, o_h, s_h\right].$$

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 $\widetilde{\mathcal{O}}\left(\frac{H}{\epsilon} \cdot \left(\max_{h\in[H]} d_{DS}^{3/2}(\mathcal{F}_h) + \log\left(\frac{1}{\delta}\right)\right)\right)$ episodes to collect $\widetilde{\mathcal{O}}\left(\frac{H}{\epsilon} \cdot \left(\max_{h\in[H]} d_{DS}^{3/2}(\mathcal{F}_h) + \log\left(\frac{1}{\delta}\right)\right)\right)$ 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 $\widetilde{\mathcal{O}}\left(\frac{H^2}{\epsilon} \cdot \left(\max_{h\in[H]} d_{DS}^{3/2}(\mathcal{F}_h) + \log\left(\frac{1}{\delta}\right)\right)\right)$ episodes we have that,

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

https://doi.org/10.52202/079017-2038

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

H Missing Details in Section 5

Proof of Theorem 5.1: Let $\pi^* \in \operatorname{argmax}_{\pi \in \Pi^L} \mathbb{E}_{s_1 \sim \mu_1}[V_1^\pi(s_1)]$, where we define $V_1^\pi(s_1) := \mathbb{E}_{o_1 \sim \mathbb{O}_1(\cdot \mid s_1), a_1 \sim \pi_1(\cdot \mid o_1)}[Q_h^\pi(z_1 = (o_1), s_1, a_1)]$. We first note the following equation

$$\begin{aligned}
&\frac{1}{T} \sum_{t \in [T]} \mathbb{E}_{s_1 \sim \mu_1}[V_1^{\pi^t}(s_1)] \\
&\quad = \mathbb{E}_{s_1 \sim \mu_1}[V_1^{\pi^*}(s_1)] + \frac{1}{T} \sum_{t \in [T]} \mathbb{E}_{s_1 \sim \mu_1}\left(\widetilde{V}_1^{\pi^t}(s_1) - V_1^{\pi^*}(s_1)\right) + \frac{1}{T} \sum_{t \in [T]} \mathbb{E}_{s_1 \sim \mu_1}\left(V_1^{\pi^t}(s_1) - \widetilde{V}_1^{\pi^t}(s_1)\right).
\end{aligned} \tag{H.1}$$

Next, we make use of the performance difference lemma [39, 1, 74] on the extended space $\prod_{h \in [H]}(\mathcal{Z}_h \times \mathcal{S})$.

Definition H.1. Consider the class of policies Π^{PL} such that at step $h \in [H]$, the policies in Π^{PL} take an action based on finite memory up to this step and the use of the underlying state, e.g., for each policy $\pi_{1:H} \in \Pi^{PL}$, $\pi_h : \mathcal{Z}_h \times \mathcal{S} \to \Delta(\mathcal{A})$.

Observation 1. Note that $\Pi^L \subseteq \Pi^{PL}$.

Lemma H.2 (Performance difference Lemma [39, 1, 74]; see e.g., Lemma 1 in [74])**.** For any pair of policies $\pi = \{\pi_h\}_{h \in [H]}, \pi' = \{\pi'_h\}_{h \in [H]} \in \Pi^{PL}$, and approximation of the Q-function of policy π, we have that for each state $s_1 \in \operatorname{supp}(\mu_1)$:

$$\begin{aligned}
&\widetilde{V}_1^\pi(z_1, s_1) - V_1^{\pi'}(z_1, s_1) \\
&= \sum_{h \in [H]} \mathbb{E}_{\overline{\tau}_h \sim \pi' \mid z_1}\left[\left\langle \widetilde{Q}_h^\pi(z_h, s_h, \cdot), \pi_h(\cdot \mid z_h, s_h) - \pi'_h(\cdot \mid z_h, s_h)\right\rangle\right] \\
&\qquad + \sum_{h \in [H]} \mathbb{E}_{\overline{\tau}_h \sim \pi' \mid z_1}\left[\widetilde{Q}_h^\pi(z_h, s_h, a_h) - \mathbb{E}_{\substack{s_{h+1} \sim \mathbb{T}_h(\cdot \mid s_h, a_h), \\ o_{h+1} \sim \mathbb{O}_{h+1}(\cdot \mid s_{h+1})}}\left[r_h(s_h, a_h) + \widetilde{V}_{h+1}^\pi(z_{h+1}, s_{h+1})\right]\right],
\end{aligned}$$

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

Setting $\pi = \pi^t \in \Pi^L \subseteq \Pi^{LP}$, and $\pi' = \pi^* \in \Pi^L \subseteq \Pi^{LP}$, where we remind that $\pi^* \in \operatorname{argmax}_{\pi \in \Pi^L} V_1^\pi(s_1)$, and for each $z_h \in \mathcal{Z}_h$ and $h \in [H]$, we abuse the notation by letting $\pi_h^t(\cdot \mid z_h, s_h) = \pi_h^t(\cdot \mid z_h)$ and $\pi_h^*(\cdot \mid z_h, s_h) = \pi_h^*(\cdot \mid z_h)$ for all $s_h \in \mathcal{S}$. The above formulation is thus simplified to

$$\begin{aligned}
&\mathbb{E}_{s_1 \sim \mu_1}[\widetilde{V}_1^{\pi^t}(s_1) - V_1^{\pi^*}(s_1)] \\
&= \sum_{h \in [H]} \mathbb{E}_{\overline{\tau}_h \sim \pi^*}\left[\left\langle \widetilde{Q}_h^{\pi^t}(z_h, s_h, \cdot), \pi_h^t(\cdot \mid z_h, s_h) - \pi_h^*(\cdot \mid z_h, s_h)\right\rangle\right] \\
&\qquad + \sum_{h \in [H]} \mathbb{E}_{\overline{\tau}_h \sim \pi^*}\left[\widetilde{Q}_h^\pi(z_h, s_h, a_h) - \mathbb{E}_{\substack{s_{h+1} \sim \mathbb{T}_h(\cdot \mid s_h, a_h), \\ o_{h+1} \sim \mathbb{O}_{h+1}(\cdot \mid s_{h+1})}}\left[r_h(s_h, a_h) + \widetilde{V}_{h+1}^\pi(z_{h+1}, s_{h+1})\right]\right] \\
&\geq \sum_{h \in [H]} \mathbb{E}_{\overline{\tau}_h \sim \pi^*}\left[\left\langle \widetilde{Q}_h^{\pi^t}(z_h, s_h, \cdot), \pi_h^t(\cdot \mid z_h, s_h) - \pi_h^*(\cdot \mid z_h, s_h)\right\rangle\right],
\end{aligned}$$

where in the inequality above we used Lemma H.3. Since our policy does not depend on the realized underlying state s_h,

$$
\begin{aligned}
&\mathbb{E}_{s_1\sim\mu_1}[\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] \\
&\qquad + \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] \\
&\qquad - \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] - 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 Lemma H.3. By averaging we get,

$$
\begin{aligned}
&\frac{1}{T}\sum_{t\in[T]} \mathbb{E}_{s_1\sim\mu_1}[\widetilde{V}_1^{\pi^t}(s_1)] \\
&\quad\geq \mathbb{E}_{s_1\sim\mu_1}[V_1^{\pi^*}(s_1)] + \frac{1}{T}\sum_{h\in[H]} \mathbb{E}_{\tau_h\sim\pi^*}\left[\sum_{t\in[T]}\left\langle \mathbb{E}_{s_h\sim\boldsymbol{b}^{\text{apx}}(z_h)}\left[\widetilde{Q}_h^{\pi^t}(z_h,s_h,\cdot)\right], \pi_h^t(\cdot\mid z_h) - \pi_h^*(\cdot\mid z_h)\right\rangle\right] \\
&\qquad - 2\cdot H\cdot \sum_{h\in[H]} \mathbb{E}_{\tau_h\sim\pi^*}\left[d_{TV}(\boldsymbol{b}_h(\tau_h), \boldsymbol{b}_h^{\text{apx}}(z_h))\right] \\
&\quad\geq \mathbb{E}_{s_1\sim\mu_1}[V_1^{\pi^*}(s_1)] + \frac{H}{T}\max_{h\in[H]} \mathbb{E}_{\tau_h\sim\pi^*}\left[\sum_{t\in[T]}\left\langle \mathbb{E}_{s_h\sim\boldsymbol{b}^{\text{apx}}(z_h)}\left[\widetilde{Q}_h^{\pi^t}(z_h,s_h,\cdot)\right], \pi_h^t(\cdot\mid z_h) - \pi_h^*(\cdot\mid z_h)\right\rangle\right] \\
&\qquad - 2\cdot H^2\cdot \max_{h\in[H]} \mathbb{E}_{\tau_h\sim\pi^*}\left[d_{TV}(\boldsymbol{b}_h(\tau_h), \boldsymbol{b}_h^{\text{apx}}(z_h))\right] \\
&\quad\geq \mathbb{E}_{s_1\sim\mu_1}[V_1^{\pi^*}(s_1)] - \frac{2H\sqrt{H\log(|\mathcal{A}|)}}{\sqrt{T}} - 2\cdot H^2\cdot \max_{h\in[H]} \mathbb{E}_{\tau_h\sim\pi^*}\left[d_{TV}(\boldsymbol{b}_h(\tau_h), \boldsymbol{b}_h^{\text{apx}}(z_h))\right],
\end{aligned}
$$

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 [10]):

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

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 estimate using Equation (H.2) in Lemma H.3, and its global near-optimality for a large enough L under γ-observability is a direct consequence of Theorem 4.1 in [26]. ■

https://doi.org/10.52202/079017-2038

Lemma H.3 (Optimistic Q-function - adapted from [48])**.** Given a policy $\pi \in \Pi^L$, and a parameter $M \in \mathbb{N}$, let $\{\widetilde{Q}_h^\pi : \mathcal{Z}_h \times \mathcal{S} \times \mathcal{A} \to [0, H]\}_{h\in[H]}$ be the output of Algorithm 3. Then, with probability at least $1-\delta$: $\forall z_h \in \mathcal{Z}_h, s_h \in \mathcal{S}, a_h \in \mathcal{A}$

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

$$\mathbb{E}_{s_1\sim\mu_1}[\widetilde{V}_1^\pi(s_1)] - \mathbb{E}_{s_1\sim\mu_1}[V_1^\pi(s_1)] \leq O\left(H^2 \cdot \sqrt{\frac{\max(O,S)\cdot S\cdot A}{M} \cdot \log\left(\frac{S\cdot A}{\delta}\right)\log\left(\frac{M\cdot S\cdot A\cdot H}{\delta}\right)}\right), \tag{H.2}$$

where $V_1^\pi(s_1) = \mathbb{E}_{o_1\sim\mathbb{O}_1(\cdot\,|\,s_1),a_1\sim\pi_1(\cdot\,|\,o_1)}[Q_h^\pi(z_1 = (o_1), s_1, a_1)]$, $\widetilde{V}_1^\pi(s_1) = \mathbb{E}_{o_1\sim\mathbb{O}_1(\cdot\,|\,s_1),a_1\sim\pi_1(\cdot\,|\,o_1)}[\widetilde{Q}_h^\pi(z_1 = (o_1), s_1, a_1)]$ and $\widetilde{V}_h^\pi(z_h, s_h) = \mathbb{E}_{a_h\sim\pi_h(\cdot\,|\,z_h)}[\widetilde{Q}_h^\pi(z_h, s_h, a_h)]$. Moreover, Algorithm 3 needs a total of $H \cdot M$ episodes from POMDP $\mathcal{P}$ and runs in time $\text{POLY}(H, M, S, A^L, O^L)$.

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:

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

Set $\delta_1 = \frac{\delta}{2\cdot S\cdot(A+1)}$. By [13], 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 $1 - \delta_1$:

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

For the rest of the proof, we condition on this event. By union bound, this happens with probability at least $1 - \frac{\delta}{2}$. We define the optimistic Q-function recursively as follows for a memory-state pair $(z_h, s_h) \in \mathcal{Z}_h \times \mathcal{S}$:

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

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

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

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

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

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

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

Thus, it holds that:

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

Thus, we conclude that

$$\begin{aligned}
&\mathbb{E}_{s_1\sim\mu_1}[\widetilde{V}^{\pi}_1(s_1)] - \mathbb{E}_{s_1\sim\mu_1}[V^{\pi}_1(s_1)] \leq \mathbb{E}_{\tau=(s_1,a_1,\ldots,s_{H+1})\sim\pi}\left[\sum_{h\in[H]} 8\cdot H\cdot C\cdot\sqrt{\frac{\max(S,O)\log(1/\delta_1)}{\max(N_h(s_h,a_h),1)}}\right]\\
&= 8\cdot H\sqrt{\max(O,S)\cdot\log(1/\delta_1)}\cdot C\cdot\sum_{h\in[H]}\mathbb{E}_{\tau=(s_1,a_1,\ldots,s_{H+1})\sim\pi}\left[\sqrt{\frac{1}{\max(N_h(s_h,a_h),1)}}\right].
\end{aligned}$$

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

Lemma H.4 (Lemma 6 in [48])**.** For each step $h \in [H]$, and state-action pair $(s_h, a_h) \in \mathcal{S}\times\mathcal{A}$, with probability at least $1-\delta_2$:

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

https://doi.org/10.52202/079017-2038

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

$$\begin{aligned}
&\mathbb{E}_{s_1 \sim \mu_1}[\widetilde{V}_1^{\pi}(s_1)] - \mathbb{E}_{s_1 \sim \mu_1}[V_1^{\pi}(s_1)] \\
&= 8\sqrt{\max(O,S) \cdot \log(1/\delta_1)} \cdot C \cdot \sum_{h \in [H]} \mathbb{E}_{\tau=(s_1,a_1,\ldots,s_{H+1}) \sim \pi} \left[\sqrt{\frac{1}{\max(N_h(s_h,a_h),1)}} \right] \\
&\leq O\left(H^2 \cdot \sqrt{\frac{\max(O,S) \cdot S \cdot A}{M}} \cdot \log\left(\frac{S \cdot A}{\delta}\right) \log\left(\frac{M \cdot S \cdot A \cdot H}{\delta}\right) \right).
\end{aligned}$$

□

Proof of Theorem 5.2: The proof of Theorem 5.2 follows by combining Theorem H.5 and Theorem H.6 below. Theorem H.5 proves that we can approximately learn a POMDP model $\mathcal{P}$ computationally and sample efficiently, thanks to the privileged information.

Theorem H.5. 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 at least $1 - \delta$, for any policy $\pi \in \Pi^{\text{gen}}$ and $h \in [H]$

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

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.11, it suffices to consider only $\pi \in \Pi_{\mathcal{S}}$ as the optimal value for policies in Π^{gen} can be achieved by those 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 $\mathcal{U}(h, \epsilon_1) = \{s_h \in \mathcal{S} \,|\, p_h(s_h) \geq \epsilon_1\}$. By the guarantee of the EULER algorithm from [89, 37], one can learn a 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 in Algorithm 4 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: for any $\pi \in \Pi_{\mathcal{S}}$

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

where the first inequality follows by [13] with probability at least $1-\delta_2$, and the second inequality uses $d_h^{\Psi(h,s_h)}(s_h) \geq \frac{p_h(s_h)}{2}$ for all $s_h \in \mathcal{U}(h,\epsilon_1)$. Similarly, for any $\pi \in \Pi_{\mathcal{S}}$

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

where similarly the first inequality follows by [13] with probability at least $1-\delta_2$, and the second inequality uses $d_h^{\Psi(h,s_h)}(s_h) \geq \frac{p_h(s_h)}{2}$ for all $s_h \in \mathcal{U}(h,\epsilon_1)$. Therefore, by a union bound, all the high probability events above hold with probability

$$1 - SH\delta_1 - SHAe^{-N\epsilon_1/8} - 2SAH\delta_2.$$

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

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

which completes the proof. □

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

Theorem H.6. Consider a γ-observable POMDP $\mathcal{P}$ (c.f. Assumption 2.2), an $\epsilon > 0$, approximate transition and emission $\{\widehat{\mathbb{T}}_h\}_{h\in[H]}$ and $\{\widehat{\mathbb{O}}_h\}_{h\in[H]}$ learned from Algorithm 4 that ensure that for any $\pi \in \Pi^{\text{gen}}$ and $h \in [H]$:

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

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

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

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 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 address this issue, our key technique is to *redirect* the transition probability into states that cannot be explored sufficiently to some highly visited states, in a new POMDP that is close to the original one in generating the beliefs.

Specifically, first, for any $\epsilon_1 > 0$, we define

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

https://doi.org/10.52202/079017-2038

By 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 $p_h(s_h) := \max_{\pi \in \Pi_S} d_h^\pi(s_h)$. 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

$$\mathbb{T}_h^{\text{trunc}}(s_{h+1} \,|\, s_h, a_h) := \mathbb{T}_h(s_{h+1} \,|\, s_h, a_h) + \frac{\sum_{s'_{h+1} \in \mathcal{S}_{h+1}^{\text{low}}} \mathbb{T}_h(s'_{h+1} \,|\, s_h, a_h)}{|\mathcal{S}_{h+1}^{\text{high}}|}, \quad \forall s_h \in \mathcal{S}, s_{h+1} \in \mathcal{S}_{h+1}^{\text{high}}, a_h \in \mathcal{A},$$

$$\mathbb{T}_h^{\text{trunc}}(s_{h+1} \,|\, s_h, a_h) := 0, \quad \forall s_h \in \mathcal{S}, s_{h+1} \in \mathcal{S}_{h+1}^{\text{low}}, a_h \in \mathcal{A}.$$

Meanwhile, for the initial state distribution, we define

$$\mu_1^{\text{trunc}}(s_1) := \mu_1(s_1) + \frac{\sum_{s'_1 \in \mathcal{S}_1^{\text{low}}} \mu_1(s'_1)}{|\mathcal{S}_1^{\text{high}}|}, \quad \forall s_1 \in \mathcal{S}_1^{\text{high}},$$
$$\mu_1^{\text{trunc}}(s_1) := 0, \quad \forall s_1 \in \mathcal{S}_1^{\text{low}}.$$

For emission, we simply define

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

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

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

for any $\overline{\tau}_h \in \overline{\mathcal{T}}_h^{\text{high}} := \{(s_{1:h}, o_{1:h}, a_{1:h-1}) \,|\, s'_h \in \mathcal{S}_{h'}^{\text{high}}, \forall h' \in [h]\}$, since some rarely visited states' probability has been redirected to the highly visited ones in $\mathcal{P}^{\text{trunc}}$. Meanwhile, it holds by a union bound that for any $h \in [H]$

$$\mathbb{P}^{\pi,\mathcal{P}}(\overline{\tau}_h \notin \overline{\mathcal{T}}_h^{\text{high}}) \leq \sum_{h' \in [h]} \mathbb{P}^{\pi,\mathcal{P}}(s_{h'} \in \mathcal{S}_{h'}^{\text{low}}) \leq HS\epsilon_1.$$

Therefore, by noticing that $\mathbb{P}^{\pi,\mathcal{P}^{\text{trunc}}}(\overline{\tau}_h) = 0$ for any $\overline{\tau}_h \notin \overline{\mathcal{T}}_h^{\text{high}}$ and $h \in [H]$, the total variation distance between the trajectory distributions in $\mathcal{P}$ and $\mathcal{P}^{\text{trunc}}$ can be bounded by

$$\sum_{\overline{\tau}_h} |\mathbb{P}^{\pi,\mathcal{P}}(\overline{\tau}_h) - \mathbb{P}^{\pi,\mathcal{P}^{\text{trunc}}}(\overline{\tau}_h)| \leq 2HS\epsilon_1. \tag{H.4}$$

On the other hand, by Equation (H.9) of Lemma H.9, we have

$$\mathbb{E}_\pi^{\mathcal{P}}[\|\boldsymbol{b}_h(\tau_h) - \boldsymbol{b}_h^{\text{trunc}}(\tau_h)\|_1] \leq 2\sum_{\overline{\tau}_h} |\mathbb{P}^{\pi,\mathcal{P}}(\overline{\tau}_h) - \mathbb{P}^{\pi,\mathcal{P}^{\text{trunc}}}(\overline{\tau}_h)| \leq 4HS\epsilon_1. \tag{H.5}$$

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

$$\widehat{\mathbb{T}}_h^{\text{trunc}}(s_{h+1} \,|\, s_h, a_h) := \widehat{\mathbb{T}}_h(s_{h+1} \,|\, s_h, a_h) + \frac{\sum_{s'_{h+1} \in \mathcal{S}_{h+1}^{\text{low}}} \widehat{\mathbb{T}}_h(s'_{h+1} \,|\, s_h, a_h)}{|\mathcal{S}_{h+1}^{\text{high}}|}, \quad \forall s_h \in \mathcal{S}, s_{h+1} \in \mathcal{S}_{h+1}^{\text{high}}, a_h \in \mathcal{A},$$

$$\widehat{\mathbb{T}}_h^{\text{trunc}}(s_{h+1} \,|\, s_h, a_h) := 0, \quad \forall s_h \in \mathcal{S}, s_{h+1} \in \mathcal{S}_{h+1}^{\text{low}}, a_h \in \mathcal{A}.$$

Meanwhile, for the initial state distribution, we define

$$\widehat{\mu}_1^{\text{trunc}}(s_1) := \widehat{\mu}_1(s_1) + \frac{\sum_{s'_1 \in \mathcal{S}_1^{\text{low}}} \widehat{\mu}_1(s'_1)}{|\mathcal{S}_1^{\text{high}}|}, \quad \forall s_1 \in \mathcal{S}_1^{\text{high}},$$
$$\widehat{\mu}_1^{\text{trunc}}(s_1) := 0, \quad \forall s_1 \in \mathcal{S}_1^{\text{low}}.$$

For emission, we define

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

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

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

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 $\mathcal{S}_h^{\text{low}}$ from the state space of $\widehat{\mathcal{P}}^{\text{trunc}}$ at each step h, which does not affect the trajectory distribution as they were not reachable in $\widehat{\mathcal{P}}^{\text{trunc}}$. 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 a $\gamma/2$-observable POMDP. Therefore, for policy class with finite memory Π^L with $L \ge \widetilde{\Omega}(\gamma^{-4} \log(S/\epsilon))$, by Theorem 4.1 in [25], 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 $\widehat{\boldsymbol{b}}_h^{\text{sub}}(\tau_h), \widehat{\boldsymbol{b}}_h'^{,\text{sub}}(z_h) \in \Delta(\mathcal{S}_h^{\text{high}})$. Now we claim that

$$\mathbb{E}_\pi^{\widehat{\mathcal{P}}^{\text{trunc}}} \|\widehat{\boldsymbol{b}}_h^{\text{trunc}}(\tau_h) - \boldsymbol{b}_h^{\text{apx}}(z_h)\|_1 \le \epsilon, \tag{H.6}$$

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 $\mathcal{S}_h^{\text{low}}$. 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 $s_h \in \mathcal{S}_h^{\text{low}}$, $\widehat{\boldsymbol{b}}_h^{\text{trunc}}(\tau_h)(s_h) = \widehat{\boldsymbol{b}}_h^{\text{sub}}(\tau_h)(s_h)$ for $s_h \in \mathcal{S}_h^{\text{high}}$.

For the total variation distance between the trajectory distributions in $\mathcal{P}^{\text{trunc}}$ and $\widehat{\mathcal{P}}^{\text{trunc}}$, it holds that by Lemma H.9

$$\begin{aligned}
&\sum_{\overline{\tau}_H} |\mathbb{P}^{\pi, \mathcal{P}^{\text{trunc}}}(\overline{\tau}_H) - \mathbb{P}^{\pi, \widehat{\mathcal{P}}^{\text{trunc}}}(\overline{\tau}_H)| \\
&\le \mathbb{E}_\pi^{\mathcal{P}^{\text{trunc}}} \left[\sum_{h \in [H]} \|\mathbb{T}_h^{\text{trunc}}(\cdot \mid s_h, a_h) - \widehat{\mathbb{T}}_h^{\text{trunc}}(\cdot \mid s_h, a_h)\|_1 + \|\mathbb{O}_h^{\text{trunc}}(\cdot \mid s_h) - \widehat{\mathbb{O}}_h^{\text{trunc}}(\cdot \mid s_h)\|_1 \right] \\
&\le \sum_{h \in [H]} \mathbb{E}_\pi^{\mathcal{P}^{\text{trunc}}} \left[\|\mathbb{T}_h(\cdot \mid s_h, a_h) - \widehat{\mathbb{T}}_h(\cdot \mid s_h, a_h)\|_1 + \|\mathbb{O}_h(\cdot \mid s_h) - \widehat{\mathbb{O}}_h(\cdot \mid s_h)\|_1 \right],
\end{aligned}$$

where the last step is by Lemma H.7.

Now by Equation (H.4), we have

$$\begin{aligned}
&\sum_h \mathbb{E}_\pi^{\mathcal{P}^{\text{trunc}}} \left[\|\mathbb{T}_h(\cdot \mid s_h, a_h) - \widehat{\mathbb{T}}_h(\cdot \mid s_h, a_h)\|_1 + \|\mathbb{O}_h(\cdot \mid s_h) - \widehat{\mathbb{O}}_h(\cdot \mid s_h)\|_1 \right] \\
&= \sum_h 8HS\epsilon_1 + \mathbb{E}_\pi^{\mathcal{P}} \left[\|\mathbb{T}_h(\cdot \mid s_h, a_h) - \widehat{\mathbb{T}}_h(\cdot \mid s_h, a_h)\|_1 + \|\mathbb{O}_h(\cdot \mid s_h) - \widehat{\mathbb{O}}_h(\cdot \mid s_h)\|_1 \right] \\
&\le \frac{\epsilon}{3} + 8H^2 S\epsilon_1,
\end{aligned}$$

https://doi.org/10.52202/079017-2038

where the last step is by Theorem H.5. Hence, by Lemma H.9, it holds that

$$\|\mathbb{P}^{\pi,\mathcal{P}^{\text{trunc}}} - \mathbb{P}^{\pi,\widehat{\mathcal{P}}^{\text{trunc}}}\|_1 \leq \frac{\epsilon}{3} + 8H^2 S\epsilon_1, \tag{H.7}$$

$$\mathbb{E}_{\pi}^{\mathcal{P}^{\text{trunc}}} \|\boldsymbol{b}_h^{\text{trunc}}(\tau_h) - \widehat{\boldsymbol{b}}_h^{\text{trunc}}(\tau_h)\|_1 \leq \frac{2\epsilon}{3} + 16H^2 S\epsilon_1. \tag{H.8}$$

Finally, we are ready to prove

$$\begin{aligned}
&\mathbb{E}_{\pi}^{\mathcal{P}} \|\boldsymbol{b}_h(\tau_h) - \boldsymbol{b}_h^{\text{apx}}(z_h)\|_1 \\
&\leq \mathbb{E}_{\pi}^{\mathcal{P}} \|\boldsymbol{b}_h(\tau_h) - \boldsymbol{b}_h^{\text{trunc}}(\tau_h)\|_1 + \mathbb{E}_{\pi}^{\mathcal{P}} \|\boldsymbol{b}_h^{\text{trunc}}(\tau_h) - \widehat{\boldsymbol{b}}_h^{\text{trunc}}(\tau_h)\|_1 + \mathbb{E}_{\pi}^{\mathcal{P}} \|\widehat{\boldsymbol{b}}_h^{\text{trunc}}(\tau_h) - \boldsymbol{b}_h^{\text{apx}}(z_h)\|_1 \\
&\leq \mathbb{E}_{\pi}^{\mathcal{P}} \|\boldsymbol{b}_h(\tau_h) - \boldsymbol{b}_h^{\text{trunc}}(\tau_h)\|_1 + \mathbb{E}_{\pi}^{\mathcal{P}^{\text{trunc}}} \|\boldsymbol{b}_h^{\text{trunc}}(\tau_h) - \widehat{\boldsymbol{b}}_h^{\text{trunc}}(\tau_h)\|_1 + \mathbb{E}_{\pi}^{\widehat{\mathcal{P}}^{\text{trunc}}} \|\widehat{\boldsymbol{b}}_h^{\text{trunc}}(\tau_h) - \boldsymbol{b}_h^{\text{apx}}(z_h)\|_1 \\
&\qquad + 4\|\mathbb{P}^{\pi,\mathcal{P}} - \mathbb{P}^{\pi,\mathcal{P}^{\text{trunc}}}\|_1 + 2\|\mathbb{P}^{\pi,\mathcal{P}^{\text{trunc}}} - \mathbb{P}^{\pi,\widehat{\mathcal{P}}^{\text{trunc}}}\|_1 \\
&\leq \mathcal{O}(H^2 S\epsilon_1) + \mathcal{O}(\epsilon).
\end{aligned}$$

Therefore, by setting $\epsilon_1 = \Theta\left(\frac{\epsilon}{H^2 S}\right)$, 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 $N = \widetilde{\Theta}(\frac{O\log(SH/\delta)}{\gamma^2\epsilon_1})$ in Equation (H.3), thus concluding our proof. □

We conclude the proof of Theorem 5.2 by combining Theorem H.5 and Theorem H.6. ■

H.1 Supporting Technical Lemmas

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

Lemma H.7. Fix $n > 0$ and an index set $S \subseteq [n]$. For two sequences $x_{1:n}$ and $y_{1:n}$ such that $x_i, y_i \in [0,1]$ for $i \in [n]$ and $\sum_i x_i = 1, \sum_i y_i = 1$, we define

$$\widehat{x}_i = x_i + \frac{\sum_{j\in S} x_j}{n - |S|}, \quad \forall i \notin S; \qquad\qquad \widehat{x}_i = 0, \quad \forall i \in S.$$

We define $\widehat{y}_{1:n}$ similarly. Then, it holds that

$$\sum_i |x_i - y_i| \geq \sum_i |\widehat{x}_i - \widehat{y}_i|.$$

Proof. Note that

$$\begin{aligned}
\sum_i |\widehat{x}_i - \widehat{y}_i| = \sum_{i\notin S} |\widehat{x}_i - \widehat{y}_i| &= \sum_{i\notin S} \left| x_i + \frac{\sum_{j\in S} x_j}{n-|S|} - y_i - \frac{\sum_{j\in S} y_j}{n-|S|} \right| \\
&\leq (n-|S|) \left| \frac{\sum_{j\in S} x_j}{n-|S|} - \frac{\sum_{j\in S} y_j}{n-|S|} \right| + \sum_{i\notin S} |x_i - y_i| \\
&\leq \sum_i |x_i - y_i|,
\end{aligned}$$

which completes the proof. □

Lemma H.8. 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

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

Meanwhile, for the first inequality, it holds that

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

concluding our lemma. □

Lemma H.9. 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 the definition of belief functions). It holds that for any $\pi\in\Pi^{\text{gen}}$

$$\begin{aligned}\left\|\mathbb{P}^{\pi,\mathcal{P}}-\mathbb{P}^{\pi,\widehat{\mathcal{P}}}\right\|_1 &= \sum_{\overline{\tau}_H}|\mathbb{P}^{\pi,\mathcal{P}}(\overline{\tau}_H)-\mathbb{P}^{\pi,\widehat{\mathcal{P}}}(\overline{\tau}_H)|\\ &\le \|\mu_1-\widehat{\mu}_1\|_1+\mathbb{E}_\pi^{\mathcal{P}}\sum_{h\in[H-1]}\|\mathbb{T}_h(\cdot\,|\,s_h,a_h)-\widehat{\mathbb{T}}_h(\cdot\,|\,s_h,a_h)\|_1\\ &\quad+\mathbb{E}_\pi^{\mathcal{P}}\sum_{h\in[H]}\|\mathbb{O}_h(\cdot\,|\,s_h)-\widehat{\mathbb{O}}_h(\cdot\,|\,s_h)\|_1,\\ \mathbb{E}_\pi^{\mathcal{P}}\|\boldsymbol{b}_h(\tau_h)-\widehat{\boldsymbol{b}}_h(\tau_h)\|_1 &\le 2\|\mu_1-\widehat{\mu}_1\|_1+2\mathbb{E}_\pi^{\mathcal{P}}\sum_{h\in[H-1]}\|\mathbb{T}_h(\cdot\,|\,s_h,a_h)-\widehat{\mathbb{T}}_h(\cdot\,|\,s_h,a_h)\|_1\\ &\quad+2\mathbb{E}_\pi^{\mathcal{P}}\sum_{h\in[H]}\|\mathbb{O}_h(\cdot\,|\,s_h)-\widehat{\mathbb{O}}_h(\cdot\,|\,s_h)\|_1,\end{aligned}$$

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.

Proof. The first inequality also appears in [44] and we provide a simplified proof here for completeness. By Lemma H.8, it holds that

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

https://doi.org/10.52202/079017-2038

where the last step is again from Lemma H.8. Therefore, by repeatedly unrolling the inequality, we proved the first result.

For the second result, we notice that by Lemma H.8, it holds that

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

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

$$\begin{aligned}\mathbb{E}_\pi^{\mathcal{P}} \|\boldsymbol{b}_h(\tau_h) - \widehat{\boldsymbol{b}}_h(\tau_h)\|_1 \leq& \sum_{\tau_h, s_h} |\mathbb{P}_h^{\pi,\mathcal{P}}(\tau_h, s_h) - \mathbb{P}_h^{\pi,\widehat{\mathcal{P}}}(\tau_h, s_h)| + \sum_{\tau_h} |\mathbb{P}_h^{\pi,\mathcal{P}}(\tau_h) - \mathbb{P}_h^{\pi,\widehat{\mathcal{P}}}(\tau_h)| \\ \leq& 2 \sum_{\overline{\tau}_H} |\mathbb{P}^{\pi,\mathcal{P}}(\overline{\tau}_H) - \mathbb{P}^{\pi,\widehat{\mathcal{P}}}(\overline{\tau}_H)|, \end{aligned} \tag{H.9}$$

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. □

Corollary H.10. Consider any two POSG instances $\mathcal{G}$, $\widehat{\mathcal{G}}$ that satisfy Assumption C.8 and the corresponding belief functions in the form of $\mathbb{P}^{\mathcal{G}}, \mathbb{P}^{\widehat{\mathcal{G}}} : \mathcal{C}_h \to \Delta(\mathcal{P}_h \times \mathcal{S})$ for any $h \in [H]$, respectively. It holds that for any $\pi \in \Pi^{\text{gen}}$

$$\begin{aligned}&\mathbb{E}_\pi^{\mathcal{G}} \|\mathbb{P}^{\mathcal{G}}(\cdot, \cdot \,|\, c_h) - \mathbb{P}^{\widehat{\mathcal{G}}}(\cdot, \cdot \,|\, c_h)\|_1 \\ &\quad \leq 2\|\mu_1 - \widehat{\mu}_1\|_1 + 2\mathbb{E}_\pi^{\mathcal{G}} \sum_{h \in [H-1]} \|\mathbb{T}_h(\cdot \,|\, s_h, a_h) - \widehat{\mathbb{T}}_h(\cdot \,|\, s_h, a_h)\|_1 + 2\mathbb{E}_\pi^{\mathcal{G}} \sum_{h \in [H]} \|\mathbb{O}_h(\cdot \,|\, s_h) - \widehat{\mathbb{O}}_h(\cdot \,|\, s_h)\|_1.\end{aligned}$$

Proof. For the second result, we notice that by Lemma H.8, it holds that

$$\begin{aligned}\sum_{c_h, p_h, s_h} |\mathbb{P}_h^{\pi,\mathcal{G}}(c_h, p_h, s_h) - \mathbb{P}_h^{\pi,\widehat{\mathcal{G}}}(c_h, p_h, s_h)| \geq& -\sum_{c_h} |\mathbb{P}_h^{\pi,\mathcal{P}}(c_h) - \mathbb{P}_h^{\pi,\widehat{\mathcal{G}}}(c_h)| \\ &+ \mathbb{E}_\pi^{\mathcal{G}} \sum_{s_h, p_h} |\mathbb{P}_h^{\pi,\mathcal{G}}(s_h, p_h \,|\, c_h) - \mathbb{P}_h^{\pi,\widehat{\mathcal{G}}}(s_h, p_h \,|\, c_h)|.\end{aligned}$$

Notice the fact that $\mathbb{P}_h^{\pi,\mathcal{G}}(s_h, p_h \,|\, c_h), \mathbb{P}_h^{\pi,\widehat{\mathcal{G}}}(s_h, p_h \,|\, c_h)$ do not depend on π due to Assumption C.8, we conclude that

$$\begin{aligned}\mathbb{E}_\pi^{\mathcal{P}} \|\mathbb{P}^{\mathcal{G}}(\cdot, \cdot \,|\, c_h) - \mathbb{P}^{\widehat{\mathcal{G}}}(\cdot, \cdot \,|\, c_h)\|_1 \leq& \sum_{c_h, p_h s_h} |\mathbb{P}_h^{\pi,\mathcal{G}}(c_h, p_h, s_h) - \mathbb{P}_h^{\pi,\widehat{\mathcal{G}}}(c_h, p_h, s_h)| + \sum_{c_h} |\mathbb{P}_h^{\pi,\mathcal{G}}(c_h) - \mathbb{P}_h^{\pi,\widehat{\mathcal{G}}}(c_h)| \\ \leq& 2 \sum_{\overline{\tau}_H} |\mathbb{P}^{\pi,\mathcal{G}}(\overline{\tau}_H) - \mathbb{P}^{\pi,\widehat{\mathcal{G}}}(\overline{\tau}_H)|, \end{aligned} \tag{H.10}$$

where the last step again comes from the fact that after marginalization, the total variation distance will not increase. By combining it with Lemma H.9, we proved the second result. □

Lemma H.11. 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

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

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

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

		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.

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:

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

where the first inequality comes from the inductive hypothesis. □

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 computational 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 [74]. 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

https://doi.org/10.52202/079017-2038

$Q \leftarrow (1-\alpha)Q + \alpha Q^{\text{target}}$). 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 condition of Definition 3.2, the function ϕ_h that satisfies the condition of Proposition 7.1 can be constructed recursively as follows for any reachable τ_h

$$\phi_h(\tau_h) := \psi_h(\phi_{h-1}(\tau_{h-1}), a_{h-1}, o_h),$$

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

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

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 [19] for m-step decodable POMDP. Here we prove it for completeness. For any reachable trajectory $\tau_h \in \mathcal{T}_h$, it holds by the belief update and induction that

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

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 construct a POSG by adding a dummy agent that has the observation being the exact state at each time step, 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 underlying state, and hence (c_h, p_h) reveals the underlying 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 [66], we proved our proposition. ■

Theorem J.1. Suppose the POSG $\mathcal{G}$ satisfies Definition 7.2. 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 $s_h \sim \widehat{g}_{i,h}(\cdot \mid c_h, p_{i,h})$, and then acts according to the expert $\pi_{i,h}$, where in the following discussions, we slightly abuse the notation and regard $\widehat{g}_{i,h}(c_h, p_{i,h})$ as a random variable following the distribution of $\widehat{g}_{i,h}(\cdot \mid c_h, p_{i,h})$. 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 from 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 $i \in [n]$,

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

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

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

Therefore, by noticing 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

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

We notice the following fact that

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

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

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

Therefore, we have

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

Correspondingly, it holds that

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

which implies that

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

 https://doi.org/10.52202/079017-2038

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

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

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

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

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 $\max_{m_i\in\mathcal{M}_{\mathcal{S},i}} v_i((m_i \diamond \pi_i) \odot \pi_{-i}) = \max_{m_i\in\mathcal{M}_{\mathcal{S},i}} v_i((m_i \diamond \pi_i) \odot \pi_{-i})$ since $\mathcal{M}_{\mathcal{S},i} \subseteq \mathcal{M}_i$ by the deterministic filter condition Definition 7.2 and Lemma J.2. Therefore, we conclude that

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

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 for any reward function and $i \in [n]$ 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

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

where the second inequality follows from 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.8 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 $\pi \in \Pi_{\mathcal{S}}$,

$$
\begin{aligned}
&\max_{i \in [n], u_i \in \Pi_i, j \in [n], h \in [H]} \mathbb{P}^{u_i \times \pi_{-i}, \mathcal{G}}(s_h \neq \widehat{g}_{j,h}(c_h, p_{j,h})) \\
&\qquad \le \max_{i \in [n], u_i \in \Pi_{\mathcal{S},i}} \mathbb{E}^{\mathcal{G}}_{u_i \times \pi_{-i}} \left[\sum_{h \in [H-1]} \|\mathbb{T}_h(\cdot \mid s_h, a_h) - \widehat{\mathbb{T}}_h(\cdot \mid s_h, a_h)\|_1 + \sum_{h \in [H]} \|\mathbb{O}_h(\cdot \mid s_h) - \widehat{\mathbb{O}}_h(\cdot \mid s_h)\|_1 \right],
\end{aligned}
$$

and

$$
\begin{aligned}
&\max_{i \in [n], m_i \in \mathcal{M}_i, j \in [n], h \in [H]} \mathbb{P}^{(m_i \diamond \pi_i) \odot \pi_{-i}, \mathcal{G}}(s_h \neq \widehat{g}_{j,h}(c_h, p_{j,h})) \\
&\qquad \le \max_{i \in [n], m_i \in \mathcal{M}_{\mathcal{S},i}} \mathbb{E}^{\mathcal{G}}_{(m_i \diamond \pi_i) \odot \pi_{-i}} \left[\sum_{h \in [H]} \|\mathbb{T}_h(\cdot \mid s_h, a_h) - \widehat{\mathbb{T}}_h(\cdot \mid s_h, a_h)\|_1 + \sum_{h \in [H-1]} \|\mathbb{O}_h(\cdot \mid s_h) - \widehat{\mathbb{O}}_h(\cdot \mid s_h)\|_1 \right].
\end{aligned}
$$

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

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

https://doi.org/10.52202/079017-2038

due to the condition in Definition 7.2. Meanwhile,

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

where the first inequality is by the first inequality in 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.9. Since π_{-i} is a fixed and fully-observable Markov policy, by Lemma H.11, we have

$$\begin{aligned}
&\mathbb{E}^{\mathcal{G}}_{u_i\times\pi_{-i}}\left[\sum_{h\in[H-1]}\|\mathbb{T}_h(\cdot\mid s_h,a_h)-\widehat{\mathbb{T}}_h(\cdot\mid s_h,a_h)\|_1+\sum_{h\in[H]}\|\mathbb{O}_h(\cdot\mid s_h)-\widehat{\mathbb{O}}_h(\cdot\mid s_h)\|_1\right]\\
&\quad\le\max_{u_i\in\Pi_{\mathcal{S},i}}\mathbb{E}^{\mathcal{G}}_{u_i\times\pi_{-i}}\left[\sum_{h\in[H-1]}\|\mathbb{T}_h(\cdot\mid s_h,a_h)-\widehat{\mathbb{T}}_h(\cdot\mid s_h,a_h)\|_1+\sum_{h\in[H]}\|\mathbb{O}_h(\cdot\mid s_h)-\widehat{\mathbb{O}}_h(\cdot\mid s_h)\|_1\right],
\end{aligned}$$

thus proving the first result of our lemma.

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

$$\begin{aligned}
&\mathbb{P}^{(m_i\diamond\pi_i)\odot\pi_{-i},\mathcal{G}}(s_h\neq\widehat{g}_{j,h}(c_h,p_{j,h}))\\
&\le\mathbb{E}^{\mathcal{G}}_{(m_i\diamond\pi_i)\odot\pi_{-i}}\left[\sum_{h\in[H-1]}\|\mathbb{T}_h(\cdot\mid s_h,a_h)-\widehat{\mathbb{T}}_h(\cdot\mid s_h,a_h)\|_1+\sum_{h\in[H]}\|\mathbb{O}_h(\cdot\mid s_h)-\widehat{\mathbb{O}}_h(\cdot\mid s_h)\|_1\right].
\end{aligned}$$

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

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

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

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

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

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

The following proof procedure follows similarly to that of Theorem H.5. 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 $\mathcal{U}(h,\epsilon_1)=\{s_h\in\mathcal{S}\,|\,p_h(s_h)\ge\epsilon_1\}$. By [37], one can learn a policy $\{\Psi_i(h,s_h)\}_{i\in[n]}$ with sample complexity $\widetilde{\mathcal{O}}(\frac{S^2A_iH^4}{\epsilon_1})$ such that $\max_{i\in[n]}d_h^{\Psi_i(h,s_h)\times\pi_{-i}}(s_h)\ge$

$\frac{p_h(s_h)}{2}$ 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 $u_i \in \Pi_{\mathcal{S},i}$:

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

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

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

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

$$1 - SHn\delta_1 - SHAe^{-N\epsilon_1/8} - 2SAH\delta_2.$$

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

$$SHA\left(nN + \widetilde{\Theta}\left(\frac{S^3AH^4}{\epsilon}\right)\right) = \widetilde{\Theta}\left(\frac{nS^2AHO + nS^3AH}{\epsilon^2} + \frac{S^4A^2H^5}{\epsilon}\right),$$

which completes the proof. □

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} as defined below.

Definition J.5. Fix $\pi \in \Pi_{\mathcal{S}}$. We define $\mathcal{M}_i^{\text{extended}}(\pi)$ 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 $\mathbb{T}_h^{\text{extended}}(s_{h+1}, a_{i,h+1} \,|\, s_h, a_{i,h}, a'_{i,h}) :=$

 https://doi.org/10.52202/079017-2038

$\mathbb{E}_{a_{-i,h}\sim\pi_h(\cdot\,|\,s_h,a_{i,h})}[\mathbb{T}_h(s_{h+1}\,|\,s_h,a'_{i,h},a_{-i,h})\pi_{h+1}(a_{i,h+1}\,|\,s_{h+1})]$, where we slightly abuse the notation of $\pi_h(a_{-i,h}\,|\,s_h,a_{i,h})$ and $\pi_h(a_{i,h}\,|\,s_h)$ by defining them as the posterior and marginal distributions induced by the joint distribution $\pi_h(a_h\,|\,s_h)$. Similarly, the reward is given by $r_h^{\text{extended}}(s_h,a_{i,h},a'_{i,h}) := \mathbb{E}_{a_{-i,h}\sim\pi_h(\cdot\,|\,s_h,a_{i,h})}[r_h(s_h,a'_{i,h},a_{-i,h})]$.

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 $\epsilon,\delta\in(0,1)$ and $\pi\in\Pi_{\mathcal{S}}$. Algorithm 6 can learn a decoding function $\widehat{g}$ such that

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

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 $\text{POLY}(S,A,H,O,\frac{1}{\epsilon})$.

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

We next introduce and prove several supporting lemmas used before.

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

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

with probability $1-\delta$.

Proof. Note the following holds

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

Therefore, we have

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

By the sample complexity of learning discrete distributions [13], we can learn Q such that $\sum_{x\in\mathcal{X},y\in\mathcal{Y}}|P(x,y)-Q(x,y)|\le\epsilon$ 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 $\mathcal{P}$ (or $\mathcal{G}$ in the multi-agent setting) under policy $\pi\in\Pi^{\text{gen}}$. We define for each $h\in[H]$, $(s_h,a_h,s_{h+1})\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}$

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

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

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

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 $\|\mathbb{T}_h(\cdot \mid s_h, a_h) - \widehat{\mathbb{T}}_h(\cdot \mid s_h, a_h)\|_1$ for specific k, h, s_h, a_h according to [13], 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 $\mathcal{P}$ (or $\mathcal{G}$ in the multi-agent setting) under some policy $\pi \in \Pi^{\text{gen}}$. We define for each $h \in [H]$, $(s_h, o_h) \in \mathcal{S} \times \mathcal{O}$

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

Then, with probability at least $1 - \delta$, 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 [13], and then taking union bound for all $k \in [K], h \in [H], s_h \in \mathcal{S}$. □

Now we switch to proving the guarantees for Algorithm 7.

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 [13], 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

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

in Algorithm 7, for some absolute constant $C_3 > 0$.

Before presenting our technical analysis, we define the following notation for the ease of presentation. We define the following approximate value functions for any policy $\pi \in \Pi$ in a backward way for $h \in [H]$ when given some approximate belief in the form of $\{\widehat{P}_h : \widehat{\mathcal{C}}_h \to \Delta(\mathcal{P}_h \times \mathcal{S})\}_{h\in[H]}$ as discussed in Section 7.2:

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

for each $(i, c_h) \in [n] \times \mathcal{C}_h$ and $\gamma_h \in \Gamma_h$, 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 the 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 $V_{i,h}^{\pi,\mathcal{G}}(c_h)$.

https://doi.org/10.52202/079017-2038

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

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

where

$$\epsilon_{\text{belief}} := \max_{h\in[H]} \max_{\pi\in\Pi} \mathbb{E}^{\mathcal{G}}_{\pi} \left\| \mathbb{P}^{\mathcal{G}}(\cdot,\cdot \mid c_h) - \widehat{P}_h(\cdot,\cdot \mid \widehat{c}_h) \right\|_1 .$$

Proof. It follows directly by combining Lemma 4 and Lemma 8 of [51]. □

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

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

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 $i \in [n]$, $\pi'_i \in \Pi_i$, $h \in [H]$

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

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 $h+1$, then by definition

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

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

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

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

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

Now we conclude

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

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

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

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 $\gamma_h\in\Gamma_h$ as proved above. 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,

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

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

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

where we notice $\mathbb{P}^{\mathcal{G}}(o_{h+1}\,|\,s_h,a_h) = \sum_{s_{h+1}}\mathbb{O}_{h+1}(o_{h+1}\,|\,s_{h+1})\mathbb{T}_h(s_{h+1}\,|\,s_h,a_h)$ 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

https://doi.org/10.52202/079017-2038

$\mathbb{E}_{s_{h+1}\sim\mathbb{T}_h(\cdot\,|\,s_h,a_h),o_{h+1}\sim\mathbb{O}_{h+1}(\cdot\,|\,s_{h+1})}\left[\widehat{V}_{i,h+1}^{(m_i\diamond\pi_i^k)\odot\pi_{-i}^k,\mathcal{G}}(c_{h+1})\right] \le H-h$. Therefore, we have

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

Now we conclude

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

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

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

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

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

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

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

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

where for any s_h, p_h, a_h, we have

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

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

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

Similarly, for value function, it holds that

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

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 that one can learn an $(\epsilon+H^2\epsilon_{\text{belief}})$-NE if $\mathcal{G}$ is zero-sum and $(\epsilon+H^2\epsilon_{\text{belief}})$-CCE if $\mathcal{G}$ is general-sum with sample complexity $\mathcal{O}(\frac{H^4SAO\log(SAHO/\delta)}{\epsilon^2})$ and computation complexity $\text{POLY}(S,\max_{h\in[H]}|\widehat{\mathcal{C}}_h|,\max_{h\in[H]}|\mathcal{P}_h|,H,\frac{1}{\epsilon},\log\frac{1}{\delta})$.

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

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

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

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

where we define the two Martingale difference sequences as follows

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

and the error of the belief is defined as

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

https://doi.org/10.52202/079017-2038

Since $|Z^1_{k,h}| \le H$, $|Z^2_{k,h}| \le H$, and $\epsilon_h(c^k_h) \le 2$, by Azuma-Hoeffding bound, we conclude with probability $1-3\delta$, the following holds

$$\sum_{k,h} Z^1_{k,h} \le \mathcal{O}(H\sqrt{HK\log\frac{1}{\delta}}), \qquad \sum_{k,h} Z^2_{k,h} \le \mathcal{O}(H\sqrt{HK\log\frac{1}{\delta}}),$$

$$\sum_{k,h}\epsilon_h(c^k_h) \le \sum_k \mathbb{E}^{\mathcal{G}}_{\pi^k}\left[\sum_h \epsilon_h(c_h)\right] + \mathcal{O}(\sqrt{HK\log\frac{1}{\delta}}) \le KH\epsilon_{\text{belief}} + \mathcal{O}(\sqrt{HK\log\frac{1}{\delta}}).$$

Meanwhile, by the pigeonhole principle, it holds that

$$\begin{aligned}\sum_{k,h} b^{k-1}_h(s^k_h,a^k_h) &\le H\sqrt{O\log(SAHK/\delta)}\sum_{k,h}\frac{1}{\sqrt{\max\{1,N^{k-1}_h(s^k_h,a^k_h)\}}}\\ &\le \mathcal{O}\left(H\sqrt{O\log(SAHK/\delta)}(H\sqrt{SAK}+HSA)\right).\end{aligned}$$

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

$$\sum_{k\in[K]}\max_{i\in[n]}\left(\max_{\pi'_i\in\Pi_i}\widehat{V}^{\pi'_i\times\pi_{-i},\mathcal{G}}_{i,1}(c^k_1) - \widehat{V}^{\pi,\mathcal{G}}_{i,1}(c^k_1)\right) \le KH^2\epsilon_{\text{belief}} + \mathcal{O}(H^2\sqrt{SAOK\log(SAHK/\delta)} + H^2SA\sqrt{O\log(SAHK/\delta)}).$$

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

$$\begin{aligned}&\sum_{k\in[K]}\max_{i\in[n]}\left(\max_{\pi'_i\in\Pi_i} V^{\pi'_i\times\pi^k_{-i},\mathcal{G}}_{i,1}(c^k_1) - V^{\pi^k,\mathcal{G}}_{i,1}(c^k_1)\right)\\ &\le \mathcal{O}(KH^2\epsilon_{\text{belief}} + H^2\sqrt{SAOK\log(SAHK/\delta)} + H^2SA\sqrt{O\log(SAHK/\delta)}).\end{aligned}$$

For the PAC guarantees, since we define $k^\star \in \arg\min_{i\in[n],k\in[K]} V^{\text{high},k}_{i,1}(c^k_1) - V^{\text{low},k}_{i,1}(c^k_1)$, we have

$$\begin{aligned}\text{CCE-gap}(\pi^{k^\star}) &\le \mathcal{O}(H^2\epsilon_{\text{belief}}) + \max_{i\in[n]}\left(V^{\text{high},k^\star}_{i,h}(c^{k^\star}_1) - V^{\text{low},k^\star}_{i,h}(c^{k^\star}_1)\right)\\ &\le \mathcal{O}(H^2\epsilon_{\text{belief}}) + \left(\frac{1}{K}\sum_{k\in[K]} V^{\text{high},k}_{i,h}(c^k_1) - V^{\text{low},k}_{i,h}(c^k_1)\right)\\ &\le \mathcal{O}(H^2\epsilon_{\text{belief}} + H^2\sqrt{SAO\log(SAHK/\delta)/K} + \frac{H^2SA}{K}\sqrt{O\log(SAHK/\delta)}).\end{aligned}$$

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

$$\begin{aligned}&\sum_{k\in[K]}\max_{i\in[n]}\left(\max_{m'_i\in\mathcal{M}_i} V^{(m'_i\diamond\pi^k_i)\odot\pi^k_{-i},\mathcal{G}}_{i,1}(c^k_1) - V^{\pi^k,\mathcal{G}}_{i,1}(c^k_1)\right)\\ &\le \mathcal{O}(KH^2\epsilon_{\text{belief}} + H^2\sqrt{SAOK\log(SAHK/\delta)} + H^2SA\sqrt{O\log(SAHK/\delta)}).\end{aligned}$$

Correspondingly, this implies that one can learn an $(\epsilon + H^2\epsilon_{\text{belief}})$-CE with sample complexity $\mathcal{O}(\frac{H^4SAO\log(SAHO/\delta)}{\epsilon^2})$ and computation complexity $\text{POLY}(S, \max_{h\in[H]}|\widehat{\mathcal{C}}_h|, \max_{h\in[H]}|\mathcal{P}_h|, H, \frac{1}{\epsilon}, \log\frac{1}{\delta})$

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

$$\widehat{V}^{\pi'_i\times\pi^k_{-i},\mathcal{G}}_{i,h}(c^k_h) - \widehat{V}^{\pi^k,\mathcal{G}}_{i,h}(c^k_h) \le V^{\text{high},k}_{i,h}(c^k_h) - V^{\text{low},k}_{i,h}(c^k_h),$$

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

Lemma J.17 (Adapted from Theorem H.5)**.** 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 $h \in [H]$

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

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 (see some examples of the controller set in Appendix C.3) into considerations when learning the models. Specifically, for the transition $\widehat{\mathbb{T}}_h$, what we estimate is only $\widehat{\mathbb{T}}_h(s_{h+1} \mid s_h, a_{\mathcal{I}_h,h})$ instead of $\widehat{\mathbb{T}}_h(s_{h+1} \mid s_h, a_h)$, where $\mathcal{T}_h \subseteq [n]$ is the controller set. 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.6, where we construct the model $\mathcal{G}^{\text{trunc}}$ for $\mathcal{G}$ in exactly the same way as constructing $\mathcal{P}^{\text{trunc}}$ for $\mathcal{P}$. Therefore, by Corollary H.10, we have

$$\mathbb{E}_{\pi}^{\mathcal{G}}[\|\mathbb{P}^{\mathcal{G}}(\cdot, \cdot \mid c_h) - \mathbb{P}^{\mathcal{G}^{\text{trunc}}}(\cdot, \cdot \mid c_h)\|_1] \le 2\sum_{\overline{\tau}_H} |\mathbb{P}^{\pi,\mathcal{G}}(\overline{\tau}_H) - \mathbb{P}^{\pi,\mathcal{G}^{\text{trunc}}}(\overline{\tau}_H)| \le 4HS\epsilon_1.$$

Meanwhile, we can construct $\widehat{\mathcal{G}}^{\text{trunc}}$ and $\widehat{\mathcal{G}}^{\text{sub}}$ using exactly the same way as for $\widehat{\mathcal{P}}^{\text{trunc}}$ and $\widehat{\mathcal{P}}^{\text{sub}}$, where $\widehat{\mathcal{G}}^{\text{sub}}$ is a $\gamma/2$-observable POSG.

Now, according to [51], for all the examples in Appendix C.3, 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., $\widehat{C}_h \le (AO)^{C\gamma^{-4}\log\frac{SH}{\epsilon_2}}$ for some absolute constant C and $\epsilon_2 \in (0,1)$, and the corresponding approximate belief $\{\widehat{P}_h : \widehat{\mathcal{C}}_h \to \Delta(\mathcal{S} \times \mathcal{P}_h)\}_{h\in[H]}$ satisfies that

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

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.6, and the remaining steps follow from the proof of Theorem H.6. 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, $r_i : \times_{i\in[n]}(\Theta_i \times \mathcal{A}_i) \to [0,1]$ 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 own type θ_i and takes the action a_i. We define a pure strategy of an agent as $s_i \in \text{ST}_i := \{\Theta_i \to \mathcal{A}_i\}$. We define $J_i(s_i, s_{-i})$ to be the expected rewards for agent i, given the pure joint strategy (s_i, s_{-i}).

By definition, $J_i(s_i, s_{-i})$ can be evaluated as

$$J_i(s_i, s_{-i}) := \mathbb{E}_{\theta\sim\mu} r_i(\theta, s_i(\theta_i), s_{-i}(\theta_{-i})).$$

Bayesian NE. We define $\gamma^\star \in \times_{i\in[n]}\Delta(\text{ST}_i)$ is an ϵ-NE is it satisfies that

$$\mathbb{E}_{s\sim\gamma^\star} J_i(s) \ge \mathbb{E}_{s\sim\gamma^\star} J_i(s_i', s_{-i}) - \epsilon, \quad \forall i \in [n], s_i' \in \text{ST}_i.$$

Bayesian CCE. We say a distribution of joint strategies $\gamma^\star \in \Delta(\times_{i\in[n]}\text{ST}_i)$ to be a ϵ-Bayesian CCE if it satisfies

$$\mathbb{E}_{s\sim\gamma^\star} J_i(s) \ge \mathbb{E}_{s\sim\gamma^\star} J_i(s_i', s_{-i}) - \epsilon, \quad \forall i \in [n], s_i' \in \text{ST}_i.$$

https://doi.org/10.52202/079017-2038

(Agent-form) Bayesian CE. We say a distribution of joint strategies $\gamma^\star \in \Delta(\times_{i\in[n]}\mathrm{ST}_i)$ to be an ϵ-agent-form Bayesian CE if it satisfies

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

where $\mathcal{M}_i = \{\Theta_i \times \mathcal{A}_i \to \mathcal{A}_i\}$ is the space for strategy modification, where m_i modifies s_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 computational complexity, e.g., [29, 32, 23].

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, *expert distillation* [14, 64, 58] and *asymmetric actor-critic* [68, 7, 3], which represent privileged *policy* and privileged *value* learning, respectively, with an emphasis on studying *both* the computational and sample efficiencies of the algorithms. Our results (as summarized in Table 1) showed that privileged information does improve learning efficiency in a series of 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 may be 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 (except Appendix G), and it would be interesting to extend the results to function-approximation settings to handle massively large state, action, and observation spaces in practice.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our abstract indeed accurately summarizes out paper's contribution and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We clearly outline all of our assumption needed to derive our theoretical results.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

https://doi.org/10.52202/079017-2038

Justification: We prove every theorem or lemma in this paper.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Yes, detailed are introduced in Appendix I

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We have included detailed introductions on the algorithms and environment to reproduce our results.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: In Appendix I.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We have reported the number of repeated experiments in Appendix I and error/variance bars in Figure 2 and Table 2.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

https://doi.org/10.52202/079017-2038

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We have provides all details in Appendix I.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: It does.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We address the societal impacts of our paper in Appendix A.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: There is no such risk.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: No external codes, data, models are used.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

https://doi.org/10.52202/079017-2038

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: We do not release any new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: No such experiments were involved.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: No such approval was needed.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.