Posterior Sampling-based Online Learning for Episodic POMDPs

Dengwang Tang[†]
eBay Inc.
dwtang@umich.edu

Dongze Ye
University of Southern California
dongzeye@usc.edu

Rahul Jain[†]
University of Southern California
Google DeepMind
rahul.jain@usc.edu

Ashutosh Nayyar
University of Southern California
ashutosn@usc.edu

Pierluigi Nuzzo
University of California, Berkeley
pierluigi.nuzzo@berkeley.edu

Abstract

Learning in POMDPs is known to be significantly harder than in MDPs. In this paper, we consider the online learning problem for episodic POMDPs with unknown transition and observation models. We propose a Posterior Sampling-based reinforcement learning algorithm for POMDPs (PS4POMDP), which is much simpler compared to state-of-the-art optimism-based online learning algorithms for POMDPs and is implementable with principled approximations. We show that the Bayesian regret of the proposed algorithm scales as the square root of the number of episodes. In a general setting, the regret scales exponentially in the horizon length H, and we show that this is inevitable by providing a lower bound. However, when the POMDP is undercomplete and weakly revealing (a common assumption in the POMDP literature), we establish a polynomial Bayesian regret bound. As a corollary, we extend PS4POMDP to multi-agent POMDPs while maintaining sublinear Bayesian regret.

1 Introduction

Markov Decision Processes (MDPs) have proved to be remarkably effective at representing sequential decision-making problems. When the model is known, dynamic programming provides an elegant framework for planning algorithms for MDPs. Unfortunately, in practical problems the MDP model is often unknown. This has necessitated the development of reinforcement learning (RL) methods to deal with such settings. Such RL methods, when combined with deep learning architectures, have then yielded a number of remarkably effective deep RL algorithms, such as DQN [Mnih et al., 2015, Osband et al., 2016] and PPO [Schulman et al., 2017], for very large state and action spaces. Often, when the model is unknown, there is a need for systematic exploration to facilitate efficient learning. Online learning methods exactly provide such systematic exploration. Many online RL algorithms build on methods developed for bandit learning [Auer et al., 2002a, Russo et al., 2018]. These algorithms can broadly be classified into two categories: *optimism-based* [Lai et al., 1985, Auer et al., 2002b, 2008], and *posterior sampling (PS)-based* [Thompson, 1933, Chapelle and Li, 2011]. While there is an elegant theory for both, often PS algorithms are found to have superior empirical performance [Osband and Van Roy, 2017]. Overall, the developments over the last decade have led to a fairly well-established theory of RL for MDPs, and effective algorithms for both offline and online settings.

[†]Work done at the University of Southern California.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

However, in many practical problems, the state is not fully observable, and partially observable Markov decision processes (POMDPs) provide more accurate representations. Motivating applications include high-stakes decision-making in healthcare [Ye et al., 2025, Tomkins et al., 2021] and security [Li et al., 2025, 2026], as well as multi-agent LLM systems [Yang et al., 2025, Chen et al., 2026], to name a few. Even when the transition and observation models are known, the planning problem for POMDPs is computationally challenging, as it requires conversion to an equivalent MDP problem over the belief state space, which is a continuum [Kumar and Varaiya, 2015]. Nevertheless, the belief state is still a sufficient statistic of the history, and many approximate planning algorithms [Shani et al., 2013, Silver and Veness, 2010] are based on this observation. When the model parameters are unknown, the POMDP learning problem becomes much harder than for MDPs since the belief state itself can no longer be computed. Thus, unlike their planning counterparts, POMDP learning problems cannot be simply reformulated as belief-state-MDP learning problems. This poses significant challenges in an online learning setting when the learning agent must perform systematic exploration, and accounts for the lack of progress in designing *effective* and *implementable* online learning algorithms for POMDPs.

In this paper, we consider episodic RL problems on finite-horizon POMDPs with finite state, action, and observation spaces. The exact models of the transition and observation kernels are unknown to the learning agent. We propose a Posterior Sampling-based RL algorithm for POMDPs (PS4POMDP), which is an adaptation of the PS method used for bandit learning [Agrawal and Goyal, 2012, Russo and Van Roy, 2016] and MDP learning problems [Osband et al., 2013, Russo and Van Roy, 2013, Ouyang et al., 2017]. PS-based algorithms show superior empirical performance over optimism-based approaches in a number of bandit and MDP learning settings [Agrawal and Goyal, 2012, Osband and Van Roy, 2017, Ouyang et al., 2017, 2019, Jafarnia-Jahromi et al., 2021a,b, Jahromi et al., 2022, Kalagarla et al., 2023]. If this is any guide, we can expect a similar observation for POMDP learning. We analyze the Bayesian regret of PS4POMDP in two settings, namely, (1) the general case, where no assumption on the POMDP is imposed, and (2) the undercomplete α-weakly revealing POMDPs [Jin et al., 2020, Liu et al., 2022], which quantify the requirement that the observations must be informative to a certain degree. We show that in general POMDP learning problems, the regret is $\mathrm{poly}(S, A, O, H) \cdot \tilde{\mathcal{O}}(\sqrt{(OA)^H K})$, where K is the number of episodes and H is the horizon length. We show that the exponential dependence on H is necessary by proving an $\Omega(\sqrt{A^{H-1}K})$ lower bound of the regret. Under the assumption that the POMDP is undercomplete and α-weakly revealing, we establish an $\tilde{\mathcal{O}}\left(\alpha^{-2}H^{2.5}S^2A^{0.5}O^{1.5}\sqrt{(1+SA/O)K}\right)$ upper bound on the Bayesian regret.

Main contributions. Our main contributions are the following: (i) We introduce a posterior sampling-based online RL algorithm for POMDPs (PS4POMDP) that is *simple and implementable with principled approximations*, and yet has a provable sublinear Bayesian regret bound that scales as $\tilde{\mathcal{O}}(\sqrt{K})$. Our algorithm is modular and can leverage the extensive literature on POMDP planners [Shani et al., 2013] and posterior inference methods [Russo et al., 2018]. The regret matches the known lower bound [Chen et al., 2023] while being polynomial in the other factors under the weakly-revealing assumption. ii) We establish an $\tilde{\mathcal{O}}\left(\alpha^{-2}H^{2.5}S^2A^{0.5}O^{1.5}\sqrt{(1+SA/O)K}\right)$ upper bound on the Bayesian regret, which is the best known for a posterior sampling algorithm for POMDPs (see [Jahromi et al., 2022]) and close to similar regret bounds for optimism-based algorithms [Liu et al., 2022, Chen et al., 2023]. (iii) As a technical contribution, we introduce a novel, tighter index change lemma (see Appendix B), which may prove useful in other contexts. Furthermore, our analysis introduces a confidence set that is specific to PS-based algorithms and different from those used in optimism-based algorithms [Liu et al., 2022, Chen et al., 2023]. The corresponding Lemmas 5.1 and 5.2 are new and may support future analysis of PS-based algorithms. As a corollary, we extend PS4POMDP to multi-agent POMDPs, and characterize its performance in terms of regret bounds. We report this extension in Appendix F.

2 Related literature

Posterior sampling. Posterior Sampling (PS) is among the earliest algorithms for online decision-making problems [Thompson, 1933]. PS-based methods maintain a posterior distribution over the unknown environment, sample a model from it at each round (or episode), and act optimally with respect to the sampled model; see [Russo et al., 2018] for a comprehensive overview. This approach

enjoys strong theoretical guarantees and competitive empirical performance across a range of online RL problems, including bandits [Agrawal and Goyal, 2012, Russo and Van Roy, 2016], MDPs [Osband et al., 2013, Osband and Van Roy, 2017, Ouyang et al., 2017, 2019, Jafarnia-Jahromi et al., 2021a,b], and also constrained MDPs [Kalagarla et al., 2023]. While exact posterior inference is computationally intractable in general, [Ishfaq et al., 2023] shows that approximate PS via Langevin Monte Carlo (LMC) suffices to obtain a sublinear regret guarantee in linear MDPs. PS-based methods have also been extended to deliver efficient exploration in deep RL settings [Osband et al., 2016, Ishfaq et al., 2023, Sasso et al., 2023] and, more recently, Large Language Model (LLM) agents [Arumugam and Griffiths, 2025].

We also note that Jahromi et al. [2022] applied PS to infinite horizon POMDPs with a known observation model. They established an $\mathcal{O}(\log T)$ instance-dependent regret bound in the finite parameter case under certain assumptions, where T is the number of online interactions. In the general case, they established an $\tilde{\mathcal{O}}(T^{2/3})$ regret bound assuming the existence of a consistent transition kernel estimator with a specific convergence rate. However, finding such an estimator is often one of the key difficulties in POMDP learning problems. Moreover, it requires keeping track of a certain hard-to-compute "pseudo-count" for every state-action pair, and makes other assumptions that are complicated and unverifiable.

POMDP learning. Our work is closest to [Jin et al., 2020, Zhan et al., 2022, Liu et al., 2022, 2023, Zhong et al., 2022, Chen et al., 2022, 2023], which also consider online learning for episodic POMDPs or the more general predictive state representation models. These papers all assumed α-weakly revealing POMDPs, or its generalization. Jin et al. [2020], Zhan et al. [2022], Liu et al. [2022], and Liu et al. [2023] proposed optimism-based algorithms while Chen et al. [2022, 2023], Zhong et al. [2022] proposed and analyzed optimistic posterior sampling algorithms. However, all such algorithms are computationally impractical or non-implementable. They either require solving joint optimization problems over the parameter and the policy space in each episode, or the computation of optimal values under *all* possible parameters in order to form the optimistic posterior distribution. Furthermore, Jin et al. [2020], Zhan et al. [2022], Chen et al. [2022] only provided sample complexity results (not regret bounds) for algorithms that are not implementable. Our results improve upon the Bayesian regret bound implied by the frequentist guarantee in [Liu et al., 2022] (after averaging over the prior and normalizing to our setting) by $\Omega(H^2\sqrt{SA})$. The regret results in [Liu et al., 2023, Zhong et al., 2022] are more difficult to compare due to differences in the assumptions. Regret bounds for algorithms presented in [Chen et al., 2023] are comparable (or slightly better) but the algorithms are practically non-implementable (i.e., there is no way to implement them even without worrying about computational tractability).

3 Preliminaries

Notations. For a positive integer n, $[n] := \{1, 2, \cdots, n\}$. For two integers $t_1 \leq t_2$, we use $t_1 : t_2$ to indicate the collection of indices $\{t_1, t_1 + 1, \cdots, t_2\}$. For example, $a_{1:4}$ stands for the vector (a_1, a_2, a_3, a_4). For a finite set Ω, $\Delta(\Omega)$ is the set of probability distributions on Ω. $\mathbf{1}_{\mathcal{E}}$ is the indicator function of the event $\mathcal{E}$. $\mathbf{e}_i$ represents a unit vector where the i-th entry is 1 and all other entries are 0. The dimension of $\mathbf{e}_i$ is inferred from the context in which it is used. $\mathbf{I}$ represents an identity matrix whose dimension is also inferred from the context. For finite sets Ω_1, Ω_2, if a function g has the form $\Omega_1 \mapsto \Delta(\Omega_2)$, we write $g(\omega_2|\omega_1) := [g(\omega_1)](\omega_2)$ as if g represents a conditional probability measure. Similarly, if $g : \Omega_1 \mapsto \Omega_2$ then we write $g(\omega_2|\omega_1) := \mathbf{1}_{\{g(\omega_1)=\omega_2\}}$. $\|\mu_1 - \mu_2\|_{\mathrm{TV}}$ represents the total variation distance between probability distributions μ_1 and μ_2. For $p > 0$, $\|\cdot\|_p$ is the standard ℓ_p-norm for vectors and ℓ_p induced norm for matrices. Finally, we use $\tilde{\mathcal{O}}, \tilde{\Omega}$ to hide absolute constants and logarithmic factors of problem parameters.

The environment model. A finite POMDP is characterized by a tuple $(\mathscr{S}, \mathscr{A}, \mathscr{O}, H, b_1, T, Z, r)$, where $\mathscr{S}$ is a finite set of states with $|\mathscr{S}| = S$; $\mathscr{A}$ is a finite set of actions with $|\mathscr{A}| = A$; $\mathscr{O}$ is a finite set of observations with $|\mathscr{O}| = O$; H is the horizon length; $b_1 \in \Delta(\mathscr{S})$ is the distribution of the initial state; $T = (T_h)_{h=1}^{H-1}, T_h : \mathscr{S} \times \mathscr{A} \mapsto \Delta(\mathscr{S})$ are the transition probabilities; $Z = (Z_h)_{h=1}^{H}, Z_h : \mathscr{S} \mapsto \Delta(\mathscr{O})$ are the observation probabilities; $r = (r_h)_{h=1}^{H}, r_h : \mathscr{O} \times \mathscr{A} \mapsto [0, 1]$ are the instantaneous reward functions.[1] For each POMDP characterized by the above tuple,

[1] This ensures that rewards reveal no additional state information beyond observations and actions.

we also define the following matrices: $\mathbb{T}_{h,a} = (T_h(s'|s,a))_{s'\in\mathscr{S},s\in\mathscr{S}}$ is the $S \times S$ probability transition matrix (where the rows represent the next state) under action $a \in \mathscr{A}$ at time h; $\mathbb{Z}_h = (Z_h(o|s))_{o\in\mathscr{O},s\in\mathscr{S}}$ is the $O \times S$ observation probability matrix at time h.

A (deterministic) policy $\pi = (\pi_h)_{h=1}^H$ is a collection of mappings $\pi_h : (\mathscr{O} \times \mathscr{A})^{h-1} \times \mathscr{O} \mapsto \mathscr{A}$, where π_h is the mapping an agent uses to choose an action at time $h \in [H]$ based on action and observation history in the current episode. Let $\mathit{\Pi}$ denote the space of all deterministic policies.

A trajectory $\tau = (o_h, a_h)_{h=1}^H$ is the complete action and observation history in a single episode. Let $\mathscr{T} = (\mathscr{O} \times \mathscr{A})^H$ denote the set of trajectories. Under a policy $\pi \in \mathit{\Pi}$, the probability of a trajectory $\tau = (o_h, a_h)_{h=1}^H$ is given by $\mathbb{P}^\pi(\tau) = \pi(\tau)\mathbb{P}^-(\tau)$, where

$$\pi(\tau) := \prod_{h=1}^{H} \pi_h(a_h|\tau_{h-1}, o_h) \tag{1}$$

$$\mathbb{P}^-(\tau) := \sum_{s_{1:H}\in\mathscr{S}^H} \left[b_1(s_1) Z_H(o_H|s_H) \prod_{h=1}^{H-1} Z_h(o_h|s_h) T_h(s_{h+1}|s_h, a_h) \right] \tag{2}$$

where $\tau_h \in (\mathscr{O} \times \mathscr{A})^h$ is the partial trajectory made up with the first h observations and actions in $\tau \in \mathscr{T}$ (with τ_0 being the empty trajectory). The above representation is particularly helpful for our analysis since it separates the "policy part" from the "environment part." The environment part can also be written as:

$$\mathbb{P}^-(\tau) = \mathbf{e}_{o_H}^T \mathbb{Z}_H \mathbb{T}_{H-1,a_{H-1}} \mathrm{diag}(\mathbb{Z}_{H-1}(o_{H-1}, \cdot)) \times \cdots \times \mathbb{T}_{1,a_1} \mathrm{diag}(\mathbb{Z}_1(o_1, \cdot)) b_1$$

where $\mathrm{diag}(\mathbf{w})$ is a diagonal matrix whose main diagonal is given by the vector $\mathbf{w}$.

The expected total reward in one episode under policy $\pi \in \mathit{\Pi}$, or the value of a policy π, is given by

$$V^\pi := \sum_{o_{1:H},a_{1:H}} \left(\mathbb{P}^\pi(o_{1:H}, a_{1:H}) \sum_{h=1}^{H} r_h(o_h, a_h) \right).$$

The maximum total expected reward in one episode over all policies $\pi \in \mathit{\Pi}$, or the value of the POMDP, is defined as $V^* = \max_{\pi\in\mathit{\Pi}} V^\pi$.

Learning agent's prior knowledge. We assume that $\mathscr{S}, \mathscr{A}, \mathscr{O}, H, r$ are known to the learning agent. The quantities b_1, T and Z are (in general) unknown to the agent. We assume that b_1, T, Z are parameterized by a parameter $\theta \in \Theta$, and the learning agent knows the parameterization (i.e., the agent knows the set Θ, and what $(b_1^\theta, T^\theta, Z^\theta)$ is for each given $\theta \in \Theta$). The learning agent's prior knowledge of the true environment θ^* is modeled by a distribution[2] $\nu^1 \in \Delta(\Theta)$. In the rest of the paper, we view θ^* as a primitive random variable with distribution ν^1. We will also add a subscript θ to the quantities defined above (e.g., $\mathbb{P}_\theta^\pi(\tau), \mathbb{P}_\theta^-(\tau), V_\theta^\pi, V_\theta^*$) to signify that they are associated with the POMDP $(\mathscr{S}, \mathscr{A}, \mathscr{O}, H, b_1^\theta, T^\theta, Z^\theta, r)$.

Learning agent's interaction with the environment. At the beginning of each episode, the learning agent chooses a potentially randomized policy in $\mathit{\Pi}$ based on past trajectories and policies. More specifically, for $k \in \mathbb{N}$, let $\mathcal{D}_k := (\tau^j, \pi^j)_{j=1}^{k-1} = (o_{1:H}^j, a_{1:H}^j, \pi^j)_{j=1}^{k-1}$ denote the data which the learning agent possesses at the beginning of the k-th episode, composed of trajectories and policies in the first $k-1$ episodes. At the beginning of the k-th episode, the learning agent chooses a random policy $\pi^k \sim \phi_k(\mathcal{D}_k)$ via a mapping $\phi_k : (\mathscr{T} \times \mathit{\Pi})^{k-1} \mapsto \Delta(\mathit{\Pi})$ and applies this policy throughout the k-th episode. We refer to $\phi = (\phi_k)_{k\in\mathbb{N}}$ as a *learning algorithm*. (We use this term to distinguish it from the term *policy*, which we use exclusively for local mappings in one episode.)

Objectives. Given the prior belief $\nu^1 \in \Delta(\Theta)$, the Bayesian regret of a learning algorithm ϕ over K episodes is defined as

$$\mathrm{BReg}(\phi, K) := \mathbb{E}^\phi \left[\sum_{k=1}^{K} (V_{\theta^*}^* - V_{\theta^*}^{\pi^k}) \right] = \int_{\theta\in\Theta} \mathbb{E}_\theta^\phi \left[\sum_{k=1}^{K} (V_\theta^* - V_\theta^{\pi^k}) \right] \mathrm{d}\nu^1(\theta)$$

[2]More formally, we assume that Θ is a Borel subset of some $\mathbb{R}^d$, the prior distribution ν^1 is a Borel measure, and the parameterization mapping $\theta \mapsto (b_1^\theta, T^\theta, Z^\theta)$ is Borel measurable.

which measures the difference between the optimal expected return when one knows θ^* and the expected return achieved by the learning algorithm ϕ. The goal of the learning agent is to choose a learning algorithm with small Bayesian regret.

Proposed algorithm. We consider the Posterior Sampling-based RL algorithm for POMDPs (PS4POMDP), given by Algorithm 1. In this algorithm, the learning agent keeps a posterior belief on the true parameter θ^* through Bayesian updates at the end of each episode, i.e., at the end of episode k, after utilizing the policy $\pi^k \in \Pi$ and observing the trajectory τ^k, the agent computes $\nu^{k+1} \in \Delta(\Theta)$ via the Bayes' rule:

$$\nu^{k+1}(\theta) \propto_\theta \mathbb{P}_\theta^{\tilde{\pi}^k}(\tau^k) \cdot \nu^k(\theta), \tag{3}$$

where $\propto_\theta$ hides the normalization constant with respect to $\theta \in \Theta$. We also assume that the agent has access to a POMDP planner that returns an optimal (or near-optimal) policy for a given POMDP – a common assumption in the literature [Jin et al., 2020, Liu et al., 2022, 2023, Chen et al., 2022, Jahromi et al., 2022, Xiong et al., 2022, Zhong et al., 2022].

Algorithm 1 PS4POMDP

Input: Prior $\nu^1 \in \Delta(\Theta)$; Number of episodes K
for $k = 1$ to K **do**
  Sample $\tilde{\theta}^k \sim \nu^k$
  Use the POMDP planner to obtain a policy $\tilde{\pi}^k \in \arg\max_{\pi \in \Pi}(V_{\tilde{\theta}^k}^{\pi})$
  Apply $\tilde{\pi}^k$ in the k-th episode
  Collect the trajectory τ^k and compute new posterior $\nu^{k+1} \in \Delta(\Theta)$ using (3)
end for

Assumptions on the environment. In this paper, we analyze the PS4POMDP algorithm in two different settings: 1) general POMDP without additional assumptions, and 2) undercomplete α-weakly revealing POMDPs first introduced by [Jin et al., 2020] in the context of POMDPs.

Assumption 3.1 (undercomplete and α-weakly revealing [Jin et al., 2020])**.** The observations are *undercomplete*, i.e., $O \geq S$, and the POMDP is α*-weakly revealing*, i.e., for all $\theta \in \Theta$ and all $h \in [H]$, the smallest singular value of the $O \times S$ observation probability matrix $\mathbb{Z}_h^\theta$ satisfies $\sigma_{\min}(\mathbb{Z}_h^\theta) \geq \alpha$.

Intuitively, Assumption 3.1 states that the observations must give a reasonable amount of information about the underlying state [Jin et al., 2020], ruling out the difficult classes of POMDPs where observations under certain states are virtually indistinguishable. In [Golowich et al., 2023], the authors show that an ε-optimal policy of a weakly revealing POMDP can be found with a quasipolynomial time planner.

4 Main results

In this section, we formally state our main results. The first two results concern the general case, where no assumptions are imposed on the POMDP. A proof outline can be found in Section 5.

Theorem 4.1. *For general POMDP learning problems, the Bayesian regret under the PS4POMDP algorithm satisfies*

$$\mathrm{BReg}(\phi^{\mathrm{PS4POMDP}}, K) \leq \tilde{\mathcal{O}}\left(H^2\sqrt{(S^2A + SO)(OA)^H K}\right)$$

The proof can be found in Appendix D.4.

Remark 4.2. Utilizing the fact that the trajectory probability can be separated into a product of the "policy part" and the "environment part," the episodic POMDP learning problem can be seen as a special case of linear bandit problem with dimension $d = (OA)^H$. Applying the standard result on the posterior sampling algorithm for linear bandits [Russo and Van Roy, 2016] we obtain an $\tilde{\mathcal{O}}(H\sqrt{O^{2H+1}A^H K})$ regret bound, where the additional O^H comes from the fact that $|\Pi| = \Omega(A^{O^H})$. The same regret bound can also be obtained if the LinUCB algorithm [Li et al., 2010] is applied instead [Lattimore and Szepesvári, 2020]. Theorem 4.1 presents an improvement over this naive bound by a factor of $\tilde{\Omega}\left(\sqrt{\frac{O^{H+1}}{H^2(S^2A+SO)}}\right)$.

The next result shows that, in the general case, the exponential dependence on H in the regret bound is unavoidable under any learning algorithm.

Proposition 4.3. *For any $A, H \geq 2$ and any $K \geq A^{H-1}$, there exists a POMDP learning problem with $S = O = 2$ such that the Bayesian regret for any learning algorithm ϕ satisfies*

$$\mathrm{BReg}(\phi, K) \geq \frac{1}{20}\sqrt{A^{H-1}K}.$$

The proof is available in Appendix C.

Next, we state our regret bound in the second setting, where the POMDP is assumed to be undercomplete and α-weakly revealing.

Theorem 4.4. *Under Assumption 3.1, the Bayesian regret under the PS4POMDP algorithm satisfies*

$$\mathrm{BReg}(\phi^{\mathrm{PS4POMDP}}, K) \leq \tilde{\mathcal{O}}\left(\alpha^{-2}H^{2.5}S^2A^{0.5}O^{1.5}\sqrt{(1+SA/O)K}\right)$$

The proof can be found in Appendix D.5.

Remark 4.5. Due to the complexity of solving POMDPs even with known parameters and the exponential growth of action and observation history, sometimes one may consider a restricted policy set $\tilde{\Pi} \subset \Pi$ (e.g., finite memory policies). Both Theorem 4.1 and Theorem 4.4 will continue to hold in this setting if the optimization oracle used in Algorithm 1 returns a best restricted policy, and the regret is defined with respect to the best restricted policy.

Finally, we state our regret upper bound when approximate posterior inference and planning are used in Algorithm 1.

Corollary 4.6. *Suppose that in episode k of Algorithm 1 an ϵ_k-optimal planner is used instead of the optimal planner. Suppose that $\tilde{\theta}^k$ is sampled from an approximate posterior $\hat{\nu}^k$ satisfying*

$$\mathbb{E}\left[\|\hat{\nu}^k - \nu^k\|_{\mathrm{TV}}\right] \leq \delta_k.$$

Assume that, conditioned on $\mathcal{D}_k$, the sampler and planner use randomness independent of θ^*. Then the Bayesian regret under approximate PS4POMDP, denoted by $\hat{\phi}^{\mathrm{PS4POMDP}}$, satisfies

$$\mathrm{BReg}(\hat{\phi}^{\mathrm{PS4POMDP}}, K) \leq \tilde{\mathcal{O}}\left(H^2\sqrt{(S^2A+SO)(OA)^HK}\right) + \sum_{k=1}^{K}(\epsilon_k + 2H\delta_k),$$

and if Assumption 3.1 holds,

$$\mathrm{BReg}(\hat{\phi}^{\mathrm{PS4POMDP}}, K) \leq \tilde{\mathcal{O}}\left(\alpha^{-2}H^{2.5}S^2A^{0.5}O^{1.5}\sqrt{(1+SA/O)K}\right) + \sum_{k=1}^{K}(\epsilon_k + 2H\delta_k).$$

The proof follows by bounding the posterior-sampling and confidence-set errors using total variation distance. The full proof is provided in Appendix D.6.

Our planning assumption is supported by existing results on approximate POMDP planning [Barenboim and Indelman, 2023, Lim et al., 2023]. Corollary 4.6 provides a simple bound accounting for posterior approximation errors; we leave a more detailed analysis of approximate posterior sampling based on, e.g., Langevin Monte Carlo [Mazumdar et al., 2020], for future work.

5 Proof outline

In this section we lay out the proof outline for the results stated in Section 4. The proof details are available in the Appendices.

5.1 Regret lower bound

Proposition 4.3 is motivated by the pathological POMDP instance in [Krishnamurthy et al., 2016] and [Jin et al., 2020]. In this POMDP, the first $H-1$ actions act as "rotating dials" to a "combination lock": One must enter a specific sequence in order to "unlock" at time H. The first $H-1$ observations

are completely uninformative, so that the learning agent has no way to learn if the entered sequence is correct or not until the very last step. Such a POMDP resembles a multi-armed bandit with A^{H-1} arms. Krishnamurthy et al. [2016] and Jin et al. [2020] established sample complexity lower bounds on these POMDPs. However, these results cannot be directly translated into lower bounds on the cumulative regret. In the proof of Proposition 4.3, we apply the standard change of measure technique [Auer et al., 2002b] on modified POMDPs from [Krishnamurthy et al., 2016, Jin et al., 2020].

5.2 Regret upper bound

At the center of our proofs is a confidence set $\bar{\Theta}(\mathcal{D}_k)$, which is a finite set of parameters whose log-likelihood given the data $\mathcal{D}_k$ is close to the maximum log-likelihood. The set is constructed such that with high probability, the true parameter θ^* is close to some parameter in $\bar{\Theta}(\mathcal{D}_k)$. The definition of $\bar{\Theta}(\mathcal{D}_k)$ is provided in Appendix D. The proof of both theorems will be based on Lemma 5.1, where we relate the regret bound to the quality of the confidence set, and Lemma 5.2, where we establish certain guarantees on the confidence set.

Lemma 5.1. *The Bayesian regret under the PS4POMDP algorithm can be bounded by*

$$\mathrm{BReg}(\phi^{\mathrm{PS4POMDP}}, K) \leq 2H + H\mathbb{E}\left[\sum_{k=1}^{K} \max_{\bar{\theta}\in\bar{\Theta}(\mathcal{D}_k)} \|\mathbb{P}_{\bar{\theta}}^{\tilde{\pi}^k} - \mathbb{P}_{\theta^*}^{\tilde{\pi}^k}\|_{\mathrm{TV}}\right]$$

where $\mathbb{P}_\theta^\pi$ is understood as a probability measure on $\mathscr{T}$, the set of action-observation trajectories.

The proof can be found in Appendix D.2.

On the other hand, through a martingale defined with the log-likelihood function, we prove Lemma 5.2, which provides a guarantee on the quality of the confidence set.

Lemma 5.2. *Under any learning algorithm ϕ, with probability at least $1 - \frac{1}{K}$,*

$$\max_{k\in[K]} \max_{\bar{\theta}\in\bar{\Theta}(\mathcal{D}_k)} \sum_{j=1}^{k} \|\mathbb{P}_{\bar{\theta}}^{\pi^j} - \mathbb{P}_{\theta^*}^{\pi^j}\|_{\mathrm{TV}}^2 \leq \tilde{\mathcal{O}}(HS^2A + HSO)$$

The proof of Lemma 5.2 is in Appendix D.3.

Given the above lemmas, to obtain an upper bound on the Bayesian regret, the only remaining task is to use an upper bound of

$$\sum_{j=1}^{k} \|\mathbb{P}_{\check{\theta}^k}^{\tilde{\pi}^j} - \mathbb{P}_{\theta^*}^{\tilde{\pi}^j}\|_{\mathrm{TV}}^2 \tag{4}$$

to derive an upper bound of

$$\sum_{j=1}^{K} \|\mathbb{P}_{\check{\theta}^j}^{\tilde{\pi}^j} - \mathbb{P}_{\theta^*}^{\tilde{\pi}^j}\|_{\mathrm{TV}} \tag{5}$$

where $\check{\theta}^k$ is a parameter in the confidence set $\bar{\Theta}(\mathcal{D}_k)$. The key difference in the above two expressions lies in the episode index of the "environment estimator" $\check{\theta}^k$: The former measures the difference of the *latest* estimated environment with the true environment under historical policies, while the latter measures the cumulative difference derived from the estimated environments *at the time*. To complete this task, we will make use of the following "index change" lemma, which is a corollary of the elliptical potential lemma used in the linear bandit literature (see, e.g., [Lattimore and Szepesvári, 2020]).

Lemma 5.3 (Index Change)**.** *Let $x_k, w_k \in \mathbb{R}^d$ be such that $\|w_k\|_2 \leq G_w$, $\|x_k\|_2 \leq G_x$ for all $k \in [K]$. Suppose that $\sum_{j=1}^{k}(x_j^T w_k)^2 \leq \beta$ for all $k \in [K]$. Then for any $\lambda > 0$,*

$$\sum_{k=1}^{K} |x_k^T w_k| \leq \sqrt{(\lambda+\beta)dK \log\left(1 + \frac{G_w^2 G_x^2 K}{d\lambda}\right)}.$$

The proof of the lemma can be found in Appendix B.

The proof of Theorem 4.1 follows from a direct application of Lemma 5.3 on certain suitably defined $(OA)^H$-dimensional vectors. Note that the dimension of the vectors is the reason for the exponential dependence on H.

To obtain a regret polynomial in H under Assumption 3.1, we follow the same three-step procedure by [Liu et al., 2022]. Under Assumption 3.1, we use the *projected operator distances* to both upper and lower bound total variation distances between trajectory distributions. In the first two steps, we establish relationships between (4) and (5) and projected operator distances. In the final step, we apply (a more sophisticated version of) the "index change" lemma on expressions involving projected operator distances. Since each projected operator distance can be represented by inner products of S-dimensional rather than $(OA)^H$-dimensional vectors, it allows us to obtain a better constant for the regret bound.

We conclude this section by noting an important distinction of our technique in the proof of Theorem 4.4. Instead of using Proposition 22 in [Liu et al., 2022], which is based on ℓ_1-eluder dimension, we develop a new result (Proposition B.2) based on the elliptical potential lemma. Our technical tool is arguably easier to prove and comes with a tighter guarantee, which ultimately enables us to improve the Bayesian regret bound implied by the frequentist guarantee in [Liu et al., 2022] by a factor of $\tilde{\Omega}(H^2\sqrt{SA})$. We note that the tighter regret bound is not an advantage of PS4POMDP, as our new technique can also be applied in the optimism-based algorithm of [Liu et al., 2022]. A similar technique was developed in [Chen et al., 2022, 2023] independently.

6 Experiments

In this section, we show a proof-of-concept implementation of PS4POMDP.[3] To the best of our knowledge, this is the first implementation of a provably efficient POMDP learning algorithm with general, history-dependent policies.[4] Although POMDP planning is PSPACE-complete [Papadimitriou and Tsitsiklis, 1987] and exact posterior inference is intractable in general, approximate methods for both sub-problems are well-studied and have high-quality off-the-shelf implementations [Phan et al., 2019, Kurniawati et al., 2009]. By contrast, OMLE [Liu et al., 2022, 2023] requires joint model-policy optimization, for which we are not aware of a general-purpose approximation with a corresponding guarantee.

6.1 Experimental setup

We evaluate PS4POMDP under approximate posterior inference via Markov Chain Monte Carlo (MCMC) methods, namely, NUTS [Hoffman et al., 2014] and Underdamped Langevin Monte Carlo (ULMC) [Chen et al., 2014, Ishfaq et al., 2024]. Given a sampled model-parameter $\tilde{\theta}^k$, we apply SARSOP [Kurniawati et al., 2009] to compute a near-optimal policy $\tilde{\pi}^k$. We report the policy values $V_\theta^{\tilde{\pi}^k}$ and the *frequentist* cumulative expected regret $\mathrm{Reg}_\theta(K) = \sum_{k=1}^K (V_\theta^\star - V_\theta^{\tilde{\pi}^k})$, where $V_\theta^\star$ is approximated by SARSOP on the true model. Policy values are estimated with 100 rollouts for PPO baselines and $1{,}000$ rollouts for PS4POMDP.

Baselines. We use PPO [Schulman et al., 2017] and its memory-augmented variants [Smirnov and Gu, 2025] as reference points. Such PPO variants are standard deep RL baselines for partially observable RL problems [Pleines et al., 2025, Tao et al., 2025]. For clarity, we include results for four memory-augmented baselines: PPO-Stack4, PPO-GRU [Cho et al., 2014], PPO-TrXL [Dai et al., 2019], and PPO-Mamba [Gu and Dao, 2024]. PPO inputs are one-hot encoding of the observations with a normalized time feature [Pardo et al., 2018]; full details are available in Appendix G.1.

Environments. We evaluate on four tabular POMDPs. Tiger [Cassandra et al., 1994] requires an information-gathering strategy. RiverSwim [Strehl and Littman, 2008, Osband et al., 2013] is a hard-exploration problem, to which we add sensor noise. Random-S10A4 shares the observation and reward structure of RiverSwim, while Sparse-S50A4 provides a larger sparse-transition setting. Full

[3] Code is available at `https://github.com/dongzeye/ps4pomdp`

[4] Azizzadenesheli et al. [2016] show experiments for a POMDP learning algorithm that is restricted to memoryless policies, i.e., actions are conditioned on the latest observation only. Their code is not publicly available for a comparison.

environment specifications can be found in Appendix G. In all four instances we provide the learning agent with b_1 in addition to $\mathscr{S}, \mathscr{A}, \mathscr{O}, H, r$; only T and Z are inferred. This is a mild simplification of the theoretical setting in Section 3, where b_1 is in general also unknown.

6.2 Learning performance

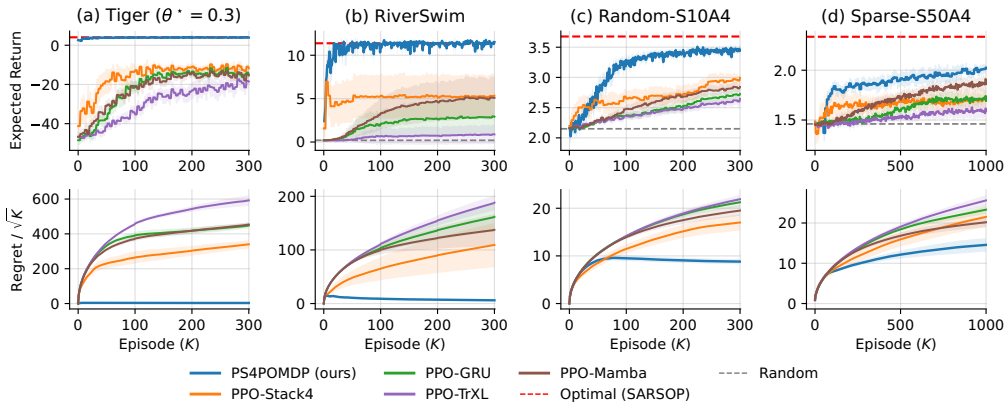


Figure 1: Experimental results on *Tiger* ($\theta^\star = 0.3$), *RiverSwim*, *Random-S10A4*, and *Sparse-S50A4*. **Top**: expected return versus the number of episodes K. **Bottom**: cumulative expected regret normalized by $\sqrt{K}$. Convergence to a constant indicates $\tilde{\mathcal{O}}(\sqrt{K})$ regret. Shaded regions denote 95% confidence intervals over independent runs.

Figure 1 presents representative experimental results across the considered domains. Despite approximation error, PS4POMDP rapidly improves its expected return and outperforms the PPO baselines in all four environments. On *Tiger*, *RiverSwim*, and *Random-S10A4*, the gap is large: PS4POMDP approaches the SARSOP reference value while the PPO variants converge to clearly suboptimal policies, and its cumulative expected regret normalized by $\sqrt{K}$ is comparatively stable. *Sparse-S50A4* is a harder scaling case. PS4POMDP remains better than the plotted PPO variants in both return and normalized regret, although the gap is narrower and its regret curve has not yet stabilized by $K = 1000$ episodes. This does not contradict the regret theory: the guarantee may involve large problem-dependent constants, particularly in larger sparse-reward POMDPs where informative trajectories are rare.

We provide detailed results per environment along with more PPO baselines in Appendices G.2 and G.3. On discretized Tiger, PS4POMDP is competitive with the best tested OMLE configuration [Liu et al., 2022]; the comparison is reported in Appendix G.2.

6.3 Effect of the planner

We test the modularity of PS4POMDP by replacing SARSOP with POMCPOW [Sunberg and Kochenderfer, 2018], DESPOT [Somani et al., 2013], or AdaOPS [Wu et al., 2021]. Table 1 reports cumulative regret over $K = 200$ learning episodes. All four planner variants achieve lower mean regret than the best PPO baseline in each environment. DESPOT and AdaOPS are close to SARSOP, while POMCPOW incurs higher regret on RiverSwim. These results support the use of online planners within PS4POMDP, while illustrating the effect of planner choice. Appendix G.6 gives the configurations and a separate comparison on the true models.

Additional results on sensitivity to posterior inference and the prior are provided in Appendices G.7 and G.8. We further discuss computational cost and report inference and planning times in Appendix G.9.

7 Discussion

We introduced an online learning algorithm for POMDPs. Although POMDP planning is computationally challenging, a number of effective approximate POMDP solvers are available when the

Table 1: Online-planner ablation: cumulative regret at $K = 200$ (lower is better), mean $\pm$ 95% CI over 10 seeds. All columns use realized learning-episode returns, unlike the separate policy evaluations in Figure 1. The reference $V^\star$ is approximated by SARSOP on the true model.

Environment	SARSOP	POMCPOW	DESPOT	AdaOPS	Best PPO
RiverSwim	$\mathbf{203.3 \pm 36.1}$	665.0 ± 45.8	214.9 ± 43.7	239.9 ± 43.3	1645.0 ± 435.2
Random-S10A4	$\mathbf{184.8 \pm 20.8}$	208.7 ± 19.8	204.3 ± 23.8	217.4 ± 15.3	276.4 ± 20.7
Sparse-S50A4	$\mathbf{133.0 \pm 12.2}$	139.7 ± 20.2	134.5 ± 13.4	138.1 ± 13.4	159.7 ± 15.0

model is known [Shani et al., 2013]. Learning when the model is unknown remains more difficult. Recent work has made substantial theoretical progress [Jin et al., 2020], but developing practical online learning methods with regret guarantees remains an important challenge.

We proposed a simple posterior sampling-based algorithm that is implementable with principled approximations by combining off-the-shelf POMDP planners and approximate posterior samplers; planning and posterior approximation errors are absorbed by an additional term in our regret bound (Corollary 4.6). Empirically, on several tabular POMDPs, the algorithm attains strong performance, outperforming PPO baselines despite the approximations involved.

Limitations and future work. More work needs to be done to make PS4POMDP scalable to large POMDP settings. Our regret guarantees are Bayesian. It would also be useful to extend PS4POMDP to continuous-state settings with function approximation. Another direction is to apply PS4POMDP or its extensions to real-world problems and study how computational limits on posterior inference and planning affect learning performance.

Acknowledgments and Disclosure of Funding

This project was partially supported by the following grants: Office of Naval Research MURI-N00014-24-1-2742, and National Science Foundation (NSF) grants 2514748, 2448886, 2025732, and 2448885.

The authors declare no competing interests.

References

Shipra Agrawal and Navin Goyal. Analysis of Thompson sampling for the multi-armed bandit problem. In *Conference on learning theory*, pages 39–1. JMLR Workshop and Conference Proceedings, 2012.

Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *Journal of machine learning research*, 15: 2773–2832, 2014.

Dilip Arumugam and Thomas L Griffiths. Toward efficient exploration by large language model agents. *arXiv preprint arXiv:2504.20997*, 2025.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47:235–256, 2002a.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002b.

Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems*, 21, 2008.

Kamyar Azizzadenesheli, Alessandro Lazaric, and Animashree Anandkumar. Reinforcement learning of POMDPs using spectral methods. In *Conference on Learning Theory*, pages 193–256. PMLR, 2016.

Moran Barenboim and Vadim Indelman. Online POMDP planning with anytime deterministic guarantees. In *Advances in Neural Information Processing Systems*, volume 36, pages 79886–79902. Curran Associates, Inc., 2023.

Alexandra Carpentier, Claire Vernade, and Yasin Abbasi-Yadkori. The elliptical potential lemma revisited. *arXiv preprint arXiv:2010.10182*, 2020.

Anthony R Cassandra, Leslie Pack Kaelbling, and Michael L Littman. Acting optimally in partially observable stochastic domains. In *Proceedings of the Twelfth AAAI National Conference on Artificial Intelligence*, pages 1023–1028, 1994.

Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. *Advances in neural information processing systems*, 24, 2011.

Fan Chen, Yu Bai, and Song Mei. Partially observable RL with B-stability: Unified structural condition and sharp sample-efficient algorithms. *arXiv preprint arXiv:2209.14990*, 2022.

Fan Chen, Huan Wang, Caiming Xiong, Song Mei, and Yu Bai. Lower bounds for learning in revealing POMDPs. *arXiv preprint arXiv:2302.01333*, 2023.

Tianqi Chen, Emily Fox, and Carlos Guestrin. Stochastic gradient Hamiltonian Monte Carlo. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1683–1691. PMLR, 2014. URL `https://proceedings.mlr.press/v32/cheni14.html`.

Yiqun Chen, Wei Yang, Erhan Zhang, Shijie Wang, Qi Liu, Zechun Niu, Bin Zhang, Haitao Li, Rui Li, Lingyong Yan, Jinyuan Feng, Biqing Qi, Xiaochi Wei, Yan Gao, Yi Wu, Yao Hu, and Jiaxin Mao. UnityMAS-O: A general RL optimization framework for LLM-based multi-agent systems. *arXiv preprint arXiv:2605.26646*, 2026. URL `https://arxiv.org/abs/2605.26646`.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G Carbonell, Quoc Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 2978–2988, 2019.

Tri Dao and Albert Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. *arXiv preprint arXiv:2405.21060*, 2024.

Noah Golowich, Ankur Moitra, and Dhruv 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.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First conference on language modeling*, 2024.

Zhaohan Daniel Guo, Shayan Doroudi, and Emma Brunskill. A PAC RL algorithm for episodic POMDPs. In *Artificial Intelligence and Statistics*, pages 510–518. PMLR, 2016.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

Matthew D Hoffman, Andrew Gelman, et al. The No-U-Turn sampler: adaptively setting path lengths in hamiltonian Monte Carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, 2014.

Daniel Hsu, Sham M Kakade, and Tong Zhang. A spectral algorithm for learning hidden Markov models. *Journal of Computer and System Sciences*, 78(5):1460–1480, 2012.

Haque Ishfaq, Qingfeng Lan, Pan Xu, A Rupam Mahmood, Doina Precup, Anima Anandkumar, and Kamyar Azizzadenesheli. Provable and practical: Efficient exploration in reinforcement learning via langevin monte carlo. *arXiv preprint arXiv:2305.18246*, 2023.

Haque Ishfaq, Yixin Tan, Yu Yang, Qingfeng Lan, Jianfeng Lu, A Rupam Mahmood, Doina Precup, and Pan Xu. More efficient randomized exploration for reinforcement learning via approximate sampling. *arXiv preprint arXiv:2406.12241*, 2024.

Mehdi Jafarnia-Jahromi, Liyu Chen, Rahul Jain, and Haipeng Luo. Online learning for stochastic shortest path model via posterior sampling. *arXiv preprint arXiv:2106.05335*, 2021a.

Mehdi Jafarnia-Jahromi, Rahul Jain, and Ashutosh Nayyar. Learning zero-sum stochastic games with posterior sampling. *arXiv preprint arXiv:2109.03396*, 2021b.

Mehdi Jafarnia Jahromi, Rahul Jain, and Ashutosh Nayyar. Online learning for unknown partially observable MDPs. In *International Conference on Artificial Intelligence and Statistics*, pages 1712–1732. PMLR, 2022.

Chi Jin, Sham Kakade, Akshay Krishnamurthy, and Qinghua Liu. Sample-efficient reinforcement learning of undercomplete POMDPs. *Advances in Neural Information Processing Systems*, 33: 18530–18539, 2020.

Krishna C Kalagarla, Rahul Jain, and Pierluigi Nuzzo. Safe posterior sampling for constrained MDPs with bounded constraint violation. *arXiv preprint arXiv:2301.11547*, 2023.

Akshay Krishnamurthy, Alekh Agarwal, and John Langford. PAC reinforcement learning with rich observations. *Advances in Neural Information Processing Systems*, 29, 2016.

Panqanamala Ramana Kumar and Pravin Varaiya. *Stochastic systems: Estimation, identification, and adaptive control*. SIAM, 2015.

Hanna Kurniawati, David Hsu, and Wee Sun Lee. Sarsop: Efficient point-based POMDP planning by approximating optimally reachable belief spaces. In *Robotics: Science and Systems*, 2009.

Tze Leung Lai, Herbert Robbins, et al. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Jonathan Lee, Alekh Agarwal, Christoph Dann, and Tong Zhang. Learning in POMDPs is sample-efficient with hindsight observability. In *International Conference on Machine Learning*, pages 18733–18773. PMLR, 2023.

David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

Jiate Li, Meng Pang, and Binghui Wang. Practicable black-box evasion attacks on link prediction in dynamic graphs—a graph sequential embedding method. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 26265–26272, 2025. doi: 10.1609/aaai.v39i25.34824. URL `https://ojs.aaai.org/index.php/AAAI/article/view/34824`.

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670, 2010.

Shawn Li, Chenxiao Yu, Zhiyu Ni, Hao Li, Charith Peris, Chaowei Xiao, and Yue Zhao. Defenses against prompt attacks learn surface heuristics. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10970–10987. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.502. URL `https://aclanthology.org/2026.acl-long.502/`.

Michael H. Lim, Tyler J. Becker, Mykel J. Kochenderfer, Claire J. Tomlin, and Zachary N. Sunberg. Optimality guarantees for particle belief approximation of POMDPs. *Journal of Artificial Intelligence Research*, 77:1591–1636, 2023.

Qinghua Liu, Alan Chung, Csaba Szepesvári, and Chi Jin. When is partially observable reinforcement learning not scary? In *Conference on Learning Theory*, pages 5175–5220. PMLR, 2022.

Qinghua Liu, Praneeth Netrapalli, Csaba Szepesvari, and Chi 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.

Eric Mazumdar, Aldo Pacchiano, Yian Ma, Michael Jordan, and Peter Bartlett. On approximate thompson sampling with Langevin algorithms. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6797–6807. PMLR, 2020.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

Ian Osband and Benjamin Van Roy. Why is posterior sampling better than optimism for reinforcement learning? In *International conference on machine learning*, pages 2701–2710. PMLR, 2017.

Ian Osband, Daniel Russo, and Benjamin Van Roy. (More) efficient reinforcement learning via posterior sampling. *Advances in Neural Information Processing Systems*, 26, 2013.

Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. Deep exploration via bootstrapped DQN. *Advances in neural information processing systems*, 29, 2016.

Yi Ouyang, Mukul Gagrani, Ashutosh Nayyar, and Rahul Jain. Learning unknown Markov decision processes: A Thompson sampling approach. *Advances in neural information processing systems*, 30, 2017.

Yi Ouyang, Mukul Gagrani, and Rahul Jain. Posterior sampling-based reinforcement learning for control of unknown linear systems. *IEEE Transactions on Automatic Control*, 65(8):3600–3607, 2019.

Christos H Papadimitriou and John N Tsitsiklis. The complexity of Markov decision processes. *Mathematics of operations research*, 12(3):441–450, 1987.

Fabio Pardo, Arash Tavakoli, Vitaly Levdik, and Petar Kormushev. Time limits in reinforcement learning. In *International Conference on Machine Learning*, pages 4045–4054. PMLR, 2018.

Emilio Parisotto, Francis Song, Jack Rae, Razvan Pascanu, Caglar Gulcehre, Siddhant Jayakumar, Max Jaderberg, Raphael Lopez Kaufman, Aidan Clark, Seb Noury, et al. Stabilizing transformers for reinforcement learning. In *International conference on machine learning*, pages 7487–7498. PMLR, 2020.

Du Phan, Neeraj Pradhan, and Martin Jankowiak. Composable effects for flexible and accelerated probabilistic programming in numpyro. *arXiv preprint arXiv:1912.11554*, 2019.

Marco Pleines, Matthias Pallasch, Frank Zimmer, and Mike Preuss. Memory gym: Towards endless tasks to benchmark memory capabilities of agents. *Journal of Machine Learning Research*, 26(6): 1–40, 2025.

Pascal Poupart and Nikos Vlassis. Model-based Bayesian reinforcement learning in partially observable domains. In *Proc Int. Symp. on Artificial Intelligence and Mathematics,*, pages 1–2, 2008.

Stephane Ross, Brahim Chaib-draa, and Joelle Pineau. Bayes-adaptive POMDPs. *Advances in neural information processing systems*, 20, 2007.

Stéphane Ross, Joelle Pineau, Brahim Chaib-draa, and Pierre Kreitmann. A Bayesian approach for learning and planning in partially observable Markov decision processes. *Journal of Machine Learning Research*, 12(5), 2011.

Daniel Russo and Benjamin Van Roy. Eluder dimension and the sample complexity of optimistic exploration. *Advances in Neural Information Processing Systems*, 26, 2013.

Daniel Russo and Benjamin Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.

Daniel Russo and Benjamin Van Roy. An information-theoretic analysis of Thompson sampling. *The Journal of Machine Learning Research*, 17(1):2442–2471, 2016.

Daniel J Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, Zheng Wen, et al. A tutorial on Thompson sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96, 2018.

Remo Sasso, Michelangelo Conserva, and Paulo Rauber. Posterior sampling for deep reinforcement learning. In *International Conference on Machine Learning*, pages 30042–30061. PMLR, 2023.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Guy Shani, Joelle Pineau, and Robert Kaplow. A survey of point-based POMDP solvers. *Auton. Agent Multi-Agent Syst.*, 27:1–51, 2013.

David Silver and Joel Veness. Monte-Carlo planning in large POMDPs. *Adv. Neural. Inf. Process. Syst.*, 23, 2010.

Ivan Smirnov and Shangding Gu. Rlbenchnet: The right network for the right reinforcement learning task. *arXiv preprint arXiv:2505.15040*, 2025.

Adhiraj Somani, Nan Ye, David Hsu, and Wee Sun Lee. DESPOT: Online POMDP planning with regularization. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

Alexander L Strehl and Michael L Littman. An analysis of model-based interval estimation for Markov decision processes. *Journal of Computer and System Sciences*, 74(8):1309–1331, 2008.

Zachary Sunberg and Mykel Kochenderfer. Online algorithms for POMDPs with continuous state, action, and observation spaces. In *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 28, pages 259–263, 2018. doi: 10.1609/icaps.v28i1.13882.

Ruo Yu Tao, Kaicheng Guo, Cameron Allen, and George Konidaris. Benchmarking partial observability in reinforcement learning with a suite of memory-improvable domains. *arXiv preprint arXiv:2508.00046*, 2025.

William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

Sabina Tomkins, Peng Liao, Predrag Klasnja, and Susan Murphy. IntelligentPooling: practical Thompson sampling for mHealth. *Machine Learning*, 110(9):2685–2727, 2021. doi: 10.1007/s10994-021-05995-8. URL https://doi.org/10.1007/s10994-021-05995-8.

Chenyang Wu, Guoyu Yang, Zongzhang Zhang, Yang Yu, Dong Li, Wulong Liu, and Jianye Hao. Adaptive online packing-guided search for POMDPs. In *Advances in Neural Information Processing Systems*, volume 34, pages 28419–28430, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/ef41d488755367316f04fc0e0e9dc9fc-Abstract.html.

Yi Xiong, Ningyuan Chen, Xuefeng Gao, and Xiang Zhou. Sublinear regret for learning POMDPs. *Production and Operations Management*, 31(9):3491–3504, 2022.

Wei Yang, Muyan Weng, Jiacheng Pang, Defu Cao, Heng Ping, Peiyu Zhang, Shixuan Li, Yue Zhao, Qiang Yang, Mengdi Wang, and Yan Liu. Toward evolutionary intelligence: LLM-based agentic systems with multi-agent reinforcement learning. *SSRN preprint*, 2025. doi: 10.2139/ssrn.5819182. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5819182.

Dongze Ye, Haipeng Luo, Carolee Winstein, and Nicolas Schweighofer. Towards AI-based precision rehabilitation via contextual model-based reinforcement learning. *Journal of NeuroEngineering and Rehabilitation*, 22:263, 2025. doi: 10.1186/s12984-025-01771-0. URL https://doi.org/10.1186/s12984-025-01771-0.

Wenhao Zhan, Masatoshi Uehara, Wen Sun, and Jason D Lee. PAC reinforcement learning for predictive state representations. *arXiv preprint arXiv:2207.05738*, 2022.

Wenqian Zhang and Haiyan Wang. Diagnostic policies optimization for chronic diseases based on POMDP model. *Healthcare*, 10(2):283, 2022. doi: 10.3390/healthcare10020283.

Han Zhong, Wei Xiong, Sirui Zheng, Liwei Wang, Zhaoran Wang, Zhuoran Yang, and Tong Zhang. GEC: A unified framework for interactive decision making in MDP, POMDP, and beyond. *arXiv preprint arXiv:2211.01962*, 2022.

A Additional related work on POMDP learning

There have been many works focusing on reinforcement learning of episodic POMDPs. In [Ross et al., 2007, Poupart and Vlassis, 2008] and [Ross et al., 2011], the authors considered Bayesian models for POMDP learning and proposed several algorithms but without sample complexity or regret guarantees. Closely related to the POMDP learning problem is the parameter estimation problem of Hidden Markov Models (HMMs), a special case of POMDPs without actions. A number of POMDP learning algorithms [Guo et al., 2016, Azizzadenesheli et al., 2016, Xiong et al., 2022] have been developed based on the spectral method (for HMMs [Hsu et al., 2012, Anandkumar et al., 2014]) but assume that both the state and observation kernels have full rank. In contrast, we only need the observation kernels to have full rank. Furthermore, [Guo et al., 2016] only provides sample complexity analysis while [Xiong et al., 2022] find a $\tilde{\mathcal{O}}(T^{2/3})$ [5] instance-dependent regret bound where T is the learning horizon, equivalent to KH. [Azizzadenesheli et al., 2016] have a regret bound that scales linearly with the inverse of the smallest singular value of the transition kernel (and hence is unsatisfactory). [Lee et al., 2023] considered POMDP learning problems with hindsight observability, where the states are revealed to the learning agent at the end of each episode. We make no such assumption.

Whether the weakly revealing assumption holds is application-dependent. For example, medical diagnostic problems can be modeled as POMDPs [Zhang and Wang, 2022]. In a model with binary disease states and diagnostic outcomes, if sensitivity plus specificity is uniformly bounded away from one, the observation matrices have smallest singular values bounded away from zero.

B Index change results

In this section we prove the two index change lemmas we used for the proof of regret upper bounds.

B.1 Proof of Lemma 5.3

We start from the following version of the elliptical potential lemma presented in [Carpentier et al., 2020].

Lemma B.1 (Elliptical Potential Lemma [Carpentier et al., 2020]). *Let $x_k \in \mathbb{R}^d$ for all $k \in [K]$ be such that $\|x_k\|_2 \leq 1$. Let $\lambda > 0$. Set $\mathbf{V}_k = \lambda \mathbf{I} + \sum_{j=1}^{k} x_j x_j^T$. Then, $\sum_{k=1}^{K} \sqrt{x_k^T \mathbf{V}_k^{-1} x_k} \leq \sqrt{dK \log\left(1 + \frac{K}{d\lambda}\right)}$.*

Through rescaling, without loss of generality, assume that $\|w_k\|_2 \leq G_w G_x$ and $\|x_k\|_2 \leq 1$. Let $\lambda > 0$. Define $\mathbf{V}_k = \frac{\lambda}{G_w^2 G_x^2}\mathbf{I} + \sum_{j=1}^{k} x_j x_j^T$. Let $\mathbf{V}_k^{1/2}$ be the symmetric square root matrix of the positive definite matrix $\mathbf{V}_k$. Using the elliptical potential lemma, we have

$$\begin{aligned}
&\sum_{k=1}^{K} |x_k^T w_k| = \sum_{k=1}^{K} |(\mathbf{V}_k^{-1/2} x_k)^T \mathbf{V}_k^{1/2} w_k| \\
\leq &\sum_{k=1}^{K} \|\mathbf{V}_k^{-1/2} x_k\|_2 \cdot \|\mathbf{V}_k^{1/2} w_k\|_2 \\
= &\sum_{k=1}^{K} \sqrt{x_k^T \mathbf{V}_k^{-1} x_k} \sqrt{w_k^T \mathbf{V}_k w_k} \\
\leq &\sqrt{dK \log\left(1 + \frac{K G_w^2 G_x^2}{d\lambda}\right)} \max_{k \in [K]} \sqrt{w_k^T \mathbf{V}_k w_k}
\end{aligned}$$

For each $k \in [K]$, we have

$$w_k^T \mathbf{V}_k w_k = \frac{\lambda}{G_x^2 G_w^2} \|w_k\|_2^2 + \sum_{j=1}^{k} (x_j^T w_k)^2 \leq \lambda + \beta.$$

[5] In this section, $\tilde{\mathcal{O}}$ or $\mathcal{O}$ hides instance-dependent constants and dependence on S, A, O, H or other related parameters.

Therefore, we conclude that

$$\sum_{k=1}^{K} |x_k^T w_k| \leq \sqrt{(\lambda+\beta)\, dK \log\left(1 + \frac{KG_w^2 G_x^2}{d\lambda}\right)}$$

for any $\lambda > 0$.

B.2 An important technical result for Theorem 4.4

The following index change result is key to improve the regret bound of [Liu et al., 2022].

Proposition B.2. *Let $w_{k,l,m}, x_{k,l,n} \in \mathbb{R}^d$ for $k \in [K], l \in [L], m \in [M], n \in [N]$. Suppose that*

$$\sum_{j=1}^{k} \left(\sum_{(l,m,n)\in[L]\times[M]\times[N]} |w_{k,l,m}^T x_{j,l,n}| \right)^2 \leq \beta,$$
$$\sum_{l=1}^{L}\sum_{m=1}^{M} \|w_{k,l,m}\|_1 \leq G_w,$$
$$\sum_{l=1}^{L}\sum_{n=1}^{N} \|x_{k,l,n}\|_1 \leq G_x,$$

for all $k \in [K]$. Then for any $\lambda > 0$,

$$\sum_{k=1}^{K} \sum_{(l,m,n)\in[L]\times[M]\times[N]} |w_{k,l,m}^T x_{k,l,n}|$$
$$\leq \sqrt{(\lambda+\beta)\, dLMK \log\left(1 + \frac{MG_w^2 G_x^2 K}{dL\lambda}\right)} \tag{6}$$

Remark B.3. The RHS of (6) is tighter than the bound of Proposition 22 by [Liu et al., 2022] (which plays a similar role in their analysis) in several ways: (1) The bound by [Liu et al., 2022] has a $\log^2(KN)$ term. In the context of its application in the main theorem, this term results in an extra $\Omega(H^2)$ term, since N is the number of partial trajectories, which is exponential in H. (2) As [Liu et al., 2022] also pointed out, the bound by [Liu et al., 2022] depends linearly on d instead of $\sqrt{d}$. (3) The bound by [Liu et al., 2022] scales linearly with $G_x G_w$, while in our bound the dependence is logarithmic.

Proof of Proposition B.2. Define $I_{k,l,m,n} := \mathrm{sgn}(w_{k,l,m}^T x_{k,l,n})$ and $\bar{x}_{k,l,m} := \sum_{n=1}^{N} x_{k,l,n} I_{k,l,m,n}$, we can then write the LHS of (6) as $\sum_{k=1}^{K} \sum_{(l,m)\in[L]\times[M]} w_{k,l,m}^T \bar{x}_{k,l,m}$.

Let $w_k = (w_{k,l,m})_{(l,m)\in[L]\times[M]}$ and $\bar{x}_k = (\bar{x}_{k,l,m})_{(l,m)\in[L]\times[M]}$ be seen as (dLM)-dimensional vectors. First, we have

$$\sum_{j=1}^{k} (w_k^T \bar{x}_j)^2$$
$$= \sum_{j=1}^{k} \left(\sum_{(l,m)\in[L]\times[M]} w_{k,l,m}^T \bar{x}_{j,l,m} \right)^2$$
$$= \sum_{j=1}^{k} \left(\sum_{(l,m,n)\in[L]\times[M]\times[N]} w_{k,l,m}^T x_{j,l,n} I_{j,l,m,n} \right)^2$$
$$\leq \sum_{j=1}^{k} \left(\sum_{(l,m,n)\in[L]\times[M]\times[N]} |w_{k,l,m}^T x_{j,l,n}| \right)^2 \leq \beta.$$

We also have

$$\|w_k\|_2 \le \|w_k\|_1 = \sum_{l=1}^{L}\sum_{m=1}^{M}\|w_{k,l,m}\|_1 \le G_w,$$
$$\|\bar{x}_k\|_2 \le \|\bar{x}_k\|_1 = \sum_{l=1}^{L}\sum_{m=1}^{M}\|\bar{x}_{k,l,m}\|_1$$
$$\le \sum_{l=1}^{L}\sum_{m=1}^{M}\sum_{n=1}^{N}\|x_{k,l,n}\|_1 \le MG_x.$$

The result is then established by applying Lemma 5.3 on the (dLM)-dimensional vectors w_k and $\bar{x}_k$. □

C Proof of Proposition 4.3

We construct the random POMDP instance as follows: Let $\epsilon \in (0, 1/2)$ be determined later and let $\mathscr{S} = \mathscr{O} = \{0, 1\}$, where 0 represents the "good" state and 1 represents the "bad" state. Let the parameter space $\Theta = \mathscr{A}^{H-1}$ be the space of $(H-1)$-sequences of actions. For each $\theta = (\theta_{1:H-1}) \in \Theta$, the final state is "good" only if the agent follows the action sequence $\theta_{1:H-1}$, i.e.,

$$b_1^\theta = \delta_0$$
$$T_h^\theta(s,a) = \begin{cases}\delta_0 & \text{if } s = 0 \text{ and } a = \theta_h \\ \delta_1 & \text{otherwise}\end{cases}$$

where $\delta_s \in \Delta(\mathscr{S})$ represents the deterministic distribution supported on state s.

The observation kernel Z^θ is the same for all $\theta \in \Theta$: Observations are uniformly random for all underlying states at all times $h \in [H-1]$. At the final timestamp, the observation kernel $\mathbb{Z}_H^\theta : \mathscr{S} \mapsto \Delta(\mathscr{O})$ is given by

$$Z_H^\theta(o|s) = \begin{cases}\frac{1}{2}+\epsilon & s = 0, o = 0 \\ \frac{1}{2}-\epsilon & s = 0, o = 1 \\ \frac{1}{2} & s = 1\end{cases}$$

The reward function is given by $r_h \equiv 0$ for $h \in [H-1]$ and $r_H(o,a) = \mathbf{1}_{\{o=0\}}$. The prior distribution ν^1 is the uniform distribution on Θ.

For the convenience of the proof, we define an additional POMDP $(b_1^{\mathbf{u}}, T^{\mathbf{u}}, Z^{\mathbf{u}})$ by

$$b_1^{\mathbf{u}} = \delta_0$$
$$T_h^{\mathbf{u}}(s,a) = \delta_0 \quad \forall s \in \mathscr{S}, a \in \mathscr{A}$$
$$Z_h^{\mathbf{u}}(o|s) = \frac{1}{2} \quad \forall o \in \mathscr{O}, s \in \mathscr{S}$$

For any policy $\pi \in \Pi$ and any trajectory $\tau = (o_{1:H}, a_{1:H}) \in \mathscr{T}$, the trajectory probability satisfies

$$\mathbb{P}_\theta^\pi(\tau) = \pi(\tau) \cdot \frac{1}{2^{H-1}}\left(\frac{1}{2} + \epsilon\iota_\theta(a_{1:H-1}, o_H)\right) \tag{7}$$
$$\mathbb{P}_{\mathbf{u}}^\pi(\tau) = \pi(\tau) \cdot \frac{1}{2^H} \tag{8}$$

where

$$\iota_\theta(a_{1:H-1}, o_H) := \begin{cases}1 & \text{if } a_{1:H-1} = \theta \text{ and } o_H = 0 \\ -1 & \text{if } a_{1:H-1} = \theta \text{ and } o_H = 1 \\ 0 & \text{if } a_{1:H-1} \neq \theta\end{cases}$$

Since the learning agent has perfect recall, the minimum Bayesian regret (equivalently, the maximum total expected reward) in K episodes can be attained through some deterministic learning algorithm ϕ. Therefore, without loss of generality, we only consider deterministic learning algorithms. More specifically, we assume that the learning algorithm is given by $\phi = (\phi_j)_{j\in\mathbb{N}}$ and ϕ_j maps the trajectories in the first $j-1$ episodes to a policy in Π. Fix ϕ, let $\mathbb{P}_\theta$ and $\mathbb{E}_\theta$ represent the probability distribution and expectation of random variables involving (potentially) multiple episodes, assuming that the underlying POMDP is given by θ, and the learning algorithm ϕ is applied. Similarly, let $\mathbb{P}_{\mathbf{u}}$ and $\mathbb{E}_{\mathbf{u}}$ represent the probability distribution and expectation of random variables under the POMDP defined by $(b_1^{\mathbf{u}}, T^{\mathbf{u}}, Z^{\mathbf{u}})$ and learning algorithm ϕ.

For each $\theta \in \Theta$, define $N_{\theta,K} := \sum_{j=1}^{K} \mathbf{1}_{\{a_{1:H-1}^j=\theta\}}$ to be a random variable denoting the number of episodes in which the learning agent's action sequence matches θ.

Lemma C.1. *For any $\theta \in \Theta$, we have*

$$\mathbb{E}_\theta[N_{\theta,K}] \le \mathbb{E}_{\mathbf{u}}[N_{\theta,K}] + \frac{K}{2}\sqrt{-\mathbb{E}_{\mathbf{u}}[N_{\theta,K}]\log(1-4\epsilon^2)}$$

Proof. First, we have

$$\begin{aligned}
&\mathbb{E}_\theta[N_{\theta,K}] - \mathbb{E}_{\mathbf{u}}[N_{\theta,K}] \\
\le &\sum_{\tilde\tau^{1:K}:\tilde\tau^j=(\tilde o_{1:H}^j,\tilde a_{1:H}^j)\in\mathscr{T}} \left[\sum_{j=1}^{K} \mathbf{1}_{\{\tilde a_{1:H-1}^j=\theta\}}\right] [\mathbb{P}_\theta(\tilde\tau^{1:K}) - \mathbb{P}_{\mathbf{u}}(\tilde\tau^{1:K})]_+ \\
\le &K\|\mathbb{P}_{\mathbf{u}}(\tau^{1:K}) - \mathbb{P}_\theta(\tau^{1:K})\|_{\mathrm{TV}}
\end{aligned}$$

By Pinsker's Inequality,

$$\|\mathbb{P}_{\mathbf{u}}(\tau^{1:K}) - \mathbb{P}_\theta(\tau^{1:K})\|_{\mathrm{TV}} \le \sqrt{\frac{1}{2}\mathbf{D}_{\mathrm{KL}}(\mathbb{P}_{\mathbf{u}}(\tau^{1:K})\|\mathbb{P}_\theta(\tau^{1:K}))}$$

where $\mathbf{D}_{\mathrm{KL}}(\mu_1\|\mu_2)$ is the KL divergence between distributions μ_1 and μ_2. Using the chain rule for the KL divergence, we have

$$\begin{aligned}
&\mathbf{D}_{\mathrm{KL}}(\mathbb{P}_{\mathbf{u}}(\tau^{1:K})\|\mathbb{P}_\theta(\tau^{1:K})) \\
=&\sum_{j=1}^{K} \mathbf{D}_{\mathrm{KL}}(\mathbb{P}_{\mathbf{u}}(\tau^j|\tau^{1:j-1})\|\mathbb{P}_\theta(\tau^j|\tau^{1:j-1})) \qquad (9)\\
=&\sum_{j=1}^{K} \sum_{\tilde\tau^{1:j-1}\in\mathscr{T}^{j-1}} \mathbb{P}_{\mathbf{u}}(\tilde\tau^{1:j-1})\mathbf{D}_{\mathrm{KL}}(\mathbb{P}_{\mathbf{u}}(\tau^j|\tilde\tau^{1:j-1})\|\mathbb{P}_\theta(\tau^j|\tilde\tau^{1:j-1}))
\end{aligned}$$

where in (9), $\mathbf{D}_{\mathrm{KL}}(\mathbb{P}_{\mathbf{u}}(\tau^j|\tau^{1:j-1})\|\mathbb{P}_\theta(\tau^j|\tau^{1:j-1}))$ is the conditional KL divergence.

For a fixed history $\tilde\tau^{1:j-1}$ before episode j, let $\tilde\pi^j = \phi_j(\tilde\tau^{1:j-1})$ be the corresponding policy the learning algorithm uses in episode j. Using (7) and (8) we have

$$\begin{aligned}
&\mathbf{D}_{\mathrm{KL}}(\mathbb{P}_{\mathbf{u}}(\tau^j|\tilde\tau^{1:j-1})\|\mathbb{P}_\theta(\tau^j|\tilde\tau^{1:j-1})) \\
=&\sum_{\tilde\tau\in\mathscr{T}} \mathbb{P}_{\mathbf{u}}^{\tilde\pi^j}(\tilde\tau)\log\left(\frac{\mathbb{P}_{\mathbf{u}}^{\tilde\pi^j}(\tilde\tau)}{\mathbb{P}_\theta^{\tilde\pi^j}(\tilde\tau)}\right) \\
=&\mathbb{P}_{\mathbf{u}}^{\tilde\pi^j}(a_{1:H-1}=\theta, o_H=0)\log\left(\frac{1}{1+2\epsilon}\right) + \mathbb{P}_{\mathbf{u}}^{\tilde\pi^j}(a_{1:H-1}=\theta, o_H=1)\log\left(\frac{1}{1-2\epsilon}\right) \\
=&\frac{1}{2}\mathbb{P}_{\mathbf{u}}^{\tilde\pi^j}(a_{1:H-1}=\theta)\left[\log\left(\frac{1}{1+2\epsilon}\right) + \log\left(\frac{1}{1-2\epsilon}\right)\right] \\
=&\frac{1}{2}\mathbb{P}_{\mathbf{u}}^{\tilde\pi^j}(a_{1:H-1}=\theta)[-\log(1-4\epsilon^2)] = \frac{1}{2}\mathbb{P}_{\mathbf{u}}(a_{1:H-1}^j=\theta|\tilde\tau^{1:j-1})[-\log(1-4\epsilon^2)].
\end{aligned}$$

Therefore,

$$\mathbf{D}_{\mathrm{KL}}(\mathbb{P}_{\mathbf{u}}(\tau^{1:K})\|\mathbb{P}_\theta(\tau^{1:K}))$$

$$= \frac{1}{2}\sum_{j=1}^{K}\sum_{\tilde{\tau}^{1:j-1}} \mathbb{P}_{\mathbf{u}}(\tilde{\tau}^{1:j-1})\mathbb{P}_{\mathbf{u}}(a^j_{1:H-1}=\theta|\tilde{\tau}^{1:j-1})\left[-\log(1-4\epsilon^2)\right]$$
$$= \frac{1}{2}\sum_{j=1}^{K}\mathbb{P}_{\mathbf{u}}(a^j_{1:H-1}=\theta)\left[-\log(1-4\epsilon^2)\right] = -\frac{1}{2}\mathbb{E}_{\mathbf{u}}[N_{\theta,K}]\log\left(1-4\epsilon^2\right)$$

and finally,

$$\mathbb{E}_\theta[N_{\theta,K}] - \mathbb{E}_{\mathbf{u}}[N_{\theta,K}] \le K\sqrt{\frac{1}{2}\mathbf{D}_{\mathrm{KL}}(\mathbb{P}_{\mathbf{u}}(\tau^{1:K})\|\mathbb{P}_\theta(\tau^{1:K}))}$$
$$= \frac{K}{2}\sqrt{-\mathbb{E}_{\mathbf{u}}[N_{\theta,K}]\log\left(1-4\epsilon^2\right)}$$

□

The conditional expected total reward $G_{\theta,K}$ under the algorithm ϕ, given that the underlying POMDP is θ, satisfies

$$G_{\theta,K} := \mathbb{E}_\theta\left[\sum_{j=1}^{K}\sum_{h=1}^{H} r_h(o^j_h, a^j_h)\right] = \sum_{j=1}^{K}\mathbb{P}_\theta(o^j_H = 0)$$
$$= \sum_{j=1}^{K}\left[\left(\frac{1}{2}+\epsilon\right)\mathbb{P}_\theta(a^j_{1:H-1}=\theta) + \frac{1}{2}\mathbb{P}_\theta(a^j_{1:H-1}\neq\theta)\right]$$
$$= \frac{K}{2} + \epsilon\sum_{j=1}^{K}\mathbb{P}_\theta(a^j_{1:H-1}=\theta) = \frac{K}{2} + \epsilon\mathbb{E}_\theta[N_{\theta,K}]$$

Define $\bar{N}_{\theta,K} := \mathbb{E}_{\mathbf{u}}[N_{\theta,K}]$ to be a fixed function of θ and K. Using Lemma C.1, we have

$$G_{\theta,K} \le \frac{K}{2} + \epsilon\left(\bar{N}_{\theta,K} + \frac{K}{2}\sqrt{-\bar{N}_{\theta,K}\log\left(1-4\epsilon^2\right)}\right)$$

If θ^* is randomly drawn from Θ following ν^1, the uniform distribution on Θ, we have

$$\mathbb{E}_{\theta^*\sim\nu^1}\left[\bar{N}_{\theta^*,K}\right] = \frac{1}{A^{H-1}}\sum_{\theta\in\Theta}\bar{N}_{\theta,K} = \frac{1}{A^{H-1}}\sum_{\theta\in\Theta}\mathbb{E}_{\mathbf{u}}[N_{\theta,K}]$$
$$= \frac{1}{A^{H-1}}\mathbb{E}_{\mathbf{u}}\left[\sum_{\theta\in\Theta}N_{\theta,K}\right] = \frac{K}{A^{H-1}},$$

and the expected total reward satisfies

$$\mathbb{E}_{\theta^*\sim\nu^1}\left[G_{\theta^*,K}\right]$$
$$\le \frac{K}{2} + \epsilon\left(\mathbb{E}_{\theta^*\sim\nu^1}[\bar{N}_{\theta^*,K}] + \frac{K}{2}\mathbb{E}_{\theta^*\sim\nu^1}\left[\sqrt{\bar{N}_{\theta^*,K}}\right]\sqrt{-\log(1-4\epsilon^2)}\right)$$
$$\le \frac{K}{2} + \epsilon\left(\mathbb{E}_{\theta^*\sim\nu^1}[\bar{N}_{\theta^*,K}] + \frac{K}{2}\sqrt{\mathbb{E}_{\theta^*\sim\nu^1}\left[\bar{N}_{\theta^*,K}\right]}\sqrt{-\log(1-4\epsilon^2)}\right)$$
$$= \frac{K}{2} + \epsilon\left(\frac{K}{A^{H-1}} + \frac{K}{2}\sqrt{\frac{K}{A^{H-1}}}\sqrt{-\log(1-4\epsilon^2)}\right)$$

For each POMDP parameterized by $\theta \in \Theta$, if one uses the open-loop policy $a_h = \theta_h$ for $h \in [H-1]$, then the expected total reward is $\frac{1}{2}+\epsilon$. This means that the optimal value satisfies $V^*_\theta \ge \frac{1}{2}+\epsilon$.

Therefore, the Bayesian regret is lower bounded by

$$\mathrm{BReg}(\phi, K) \ge \left(\frac{1}{2}+\epsilon\right)K - \mathbb{E}_{\theta^*\sim\nu^1}\left[G_{\theta^*,K}\right]$$

$$\geq \epsilon \left(K - \frac{K}{A^{H-1}} - \frac{K}{2}\sqrt{\frac{K}{A^{H-1}}}\sqrt{-\log(1-4\epsilon^2)} \right)$$

Proposition 4.3 can then be established by choosing $\epsilon = \frac{1}{4}\sqrt{\frac{A^{H-1}}{K}}$ and the inequality $-\log(1-x) \leq 4\log\left(\frac{4}{3}\right)x$ for $x \in [0, \frac{1}{4}]$ in the same way as in [Auer et al., 2002b].

D Proof of regret upper bounds

In this section, we prove both Theorem 4.1 and Theorem 4.4. Our proofs are inspired by the framework for establishing regret bounds for posterior sampling by [Russo and Van Roy, 2014] and the techniques of [Liu et al., 2022].

In this section, $\mathbb{P}_\theta^\pi$ is seen as a probability distribution on $\mathscr{T}$, the set of observable trajectories. Without loss of generality, we assume that the parameterization is direct, i.e., Θ is a subset of $[0,1]^{S+(H-1)S^2A+HSO}$, where the coordinates represent (b_1, T, Z).

While our parameter set Θ could be uncountably infinite, we construct a finite set $\bar{\Theta}$ that serves as a surrogate for Θ for technical reasons.

Proposition D.1. *For any $\epsilon > 0$ such that $\epsilon^{-1} \in \mathbb{N}$, there exists a finite set $\bar{\Theta} \subset [0,1]^{S+(H-1)S^2A+HSO}$ and a measurable function $\iota : \Theta \mapsto \bar{\Theta}$ such that*

$$\log(|\bar{\Theta}|) \leq (HS^2A + HSO)\log\left(\frac{\max\{S,O\}}{\epsilon} + 1\right)$$

and the following holds: For any $\theta \in \Theta$, $\theta' = \iota(\theta)$ satisfies

$$\|\mathbb{P}_{\theta'}^\pi - \mathbb{P}_\theta^\pi\|_{\mathrm{TV}} \leq 2H\epsilon \qquad \forall \pi \in \Pi \tag{10}$$

$$\mathbb{P}_{\theta'}^-(\tau) \geq (1+\epsilon)^{-2H}\mathbb{P}_\theta^-(\tau) \qquad \forall \tau \in \mathscr{T} \tag{11}$$

Recall that $\mathbb{P}_\theta^-$ represents the open-loop probability and is defined in (2). The proof of Proposition D.1 is presented in Appendix D.1. In the rest of the section, we set $\epsilon = 1/(2HK)$ and assume that $\bar{\Theta}$ and ι satisfy the condition of Proposition D.1. By removing points in $\bar{\Theta}$, without loss of generality, we assume that ι is surjective, i.e., for every $\bar{\theta} \in \bar{\Theta}$ there exists $\theta \in \Theta$ such that $\bar{\theta} = \iota(\theta)$.

Define the log-likelihood function

$$\ell(\theta, \mathcal{D}_k) := \sum_{j=1}^{k-1} \log(\mathbb{P}_\theta^{\pi^j}(\tau^j)) \tag{12}$$

and the maximum log-likelihood function

$$\bar{\ell}^*(\mathcal{D}_k) := \max_{\theta \in \bar{\Theta}} \ell(\theta, \mathcal{D}_k). \tag{13}$$

Note that in (13) the maximum is taken over the finite set $\bar{\Theta}$, not Θ.

Quick Fact 1: If $\bar{\theta} = \iota(\theta)$, then

$$\begin{aligned}\ell(\bar{\theta}, \mathcal{D}_k) &\geq \ell(\theta, \mathcal{D}_k) - 2(k-1)H\log(1+\epsilon) \\ &\geq \ell(\theta, \mathcal{D}_k) - 2(k-1)H\epsilon \geq \ell(\theta, \mathcal{D}_k) - 1\end{aligned}$$

Quick Fact 2: As a corollary to Quick Fact 1, for any $\theta \in \Theta$, $\ell(\theta, \mathcal{D}_k) \leq \bar{\ell}^*(\mathcal{D}_k) + 1$.

We now define the discrete confidence set

$$\bar{\Theta}(\mathcal{D}_k) := \{\bar{\theta} \in \bar{\Theta} : \ell(\bar{\theta}, \mathcal{D}_k) \geq \bar{\ell}^*(\mathcal{D}_k) - \log(K|\bar{\Theta}|) - 1\}.$$

The following lemmas form the foundation of our proof. In Lemma D.2, we bound the Bayesian regret with the quality of the confidence set in terms of the TV distance between distributions on trajectories. In Lemma D.3, we provide certain guarantees on the quality of the confidence sets.

Lemma D.2 (Lemma 5.1, restated)**.** *The Bayesian regret under the PS4POMDP algorithm can be bounded by*

$$\mathrm{BReg}(\phi^{\mathrm{PS4POMDP}}, K) \le 2H + H\mathbb{E}\left[\sum_{k=1}^{K} \max_{\bar{\theta}\in\bar{\Theta}(\mathcal{D}_k)} \|\mathbb{P}_{\bar{\theta}}^{\tilde{\pi}^k} - \mathbb{P}_{\theta^*}^{\tilde{\pi}^k}\|_{\mathrm{TV}}\right]$$

Lemma D.3 (Lemma 5.2, restated)**.** *Under any learning algorithm ϕ, with probability at least $1 - \frac{1}{K}$,*

$$\max_{k\in[K]} \max_{\bar{\theta}\in\bar{\Theta}(\mathcal{D}_k)} \sum_{j=1}^{k} \|\mathbb{P}_{\bar{\theta}}^{\pi^j} - \mathbb{P}_{\theta^*}^{\pi^j}\|_{\mathrm{TV}}^2 \le 3\log(K|\bar{\Theta}|) + 3$$

The proofs of Lemmas D.2 and D.3 can be found in Appendices D.2 and D.3.

Given the above lemmas, to obtain an upper bound on the Bayesian regret, the remaining task is purely algebraic and can be described as follows: Use an upper bound of

$$\sum_{j=1}^{k} \|\mathbb{P}_{\check{\theta}^k}^{\tilde{\pi}^j} - \mathbb{P}_{\theta^*}^{\tilde{\pi}^j}\|_{\mathrm{TV}}^2 \tag{TR-MLE}$$

to derive an upper bound of

$$\sum_{j=1}^{K} \|\mathbb{P}_{\check{\theta}^j}^{\tilde{\pi}^j} - \mathbb{P}_{\theta^*}^{\tilde{\pi}^j}\|_{\mathrm{TV}} \tag{TR-REG}$$

where $\check{\theta}^k \in \arg\max_{\bar{\theta}\in\bar{\Theta}(\mathcal{D}_k)} \|\mathbb{P}_{\bar{\theta}}^{\tilde{\pi}^k} - \mathbb{P}_{\theta^*}^{\tilde{\pi}^k}\|_{\mathrm{TV}}$.

The key difference between (TR-MLE) and (TR-REG) lies in the indexing, where the former measures the difference of the *latest* estimated environment with the true environment under historical policies, while the latter measures the cumulative differences derived from the estimated environment *at the time*.

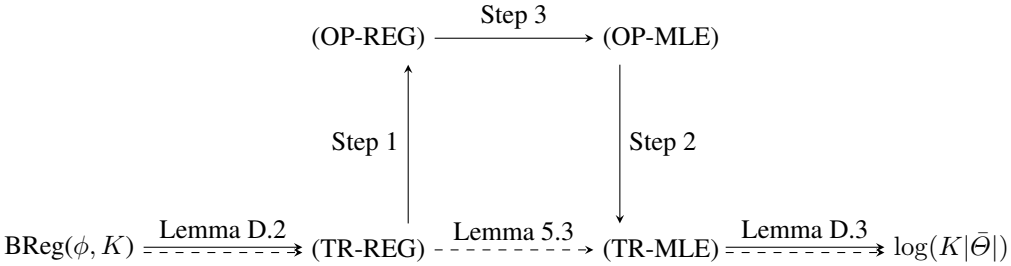


Figure 2: Proof road map of Theorem 4.1 (dashed) and Theorem 4.4 (solid). For Theorem 4.4, the bound is developed via auxiliary quantities called the projected operator distances. In this diagram, $A \to B$ means "A is upper bounded by some function of B." While "TR" stands for "trajectory-based", "OP" stands for "operator-based."

D.1 Proof of Proposition D.1

For any finite set $\mathcal{X}$, define $\bar{\Delta}^\epsilon(\mathcal{X}) \subset \Delta(\mathcal{X})$ by

$$\bar{\Delta}^\epsilon(\mathcal{X}) = \left\{ \frac{v}{\|v\|_1} : \|v\|_1 > 0,\ \frac{v_x}{\epsilon/|\mathcal{X}|} \in \mathbb{Z}_{\ge 0}, v_x \le 1 \quad \forall x \in \mathcal{X} \right\}.$$

Then we have $|\bar{\Delta}^\epsilon(\mathcal{X})| \le \left(\frac{|\mathcal{X}|}{\epsilon} + 1\right)^{|\mathcal{X}|}$. Furthermore, for any $\mu \in \Delta(\mathcal{X})$, set

$$v = \frac{1}{|\mathcal{X}|/\epsilon}\lceil(|\mathcal{X}|/\epsilon)\mu\rceil, \qquad \bar{\mu} = \frac{v}{\|v\|_1}, \tag{14}$$

we have $\|v-\mu\|_1 \le \epsilon$ and

$$\|v\|_1 \le \|\mu\|_1 + \|v-\mu\|_1 \le 1+\epsilon.$$

Therefore,

$$\begin{aligned}\|\mu-\bar{\mu}\|_{\mathrm{TV}} &\le \frac{1}{2}\|\mu-v\|_1 + \frac{1}{2}\|v-\bar{\mu}\|_1 \le \frac{1}{2}\epsilon + \frac{1}{2}\left(\|v\|_1 - 1\right) \le \epsilon \\ \bar{\mu} &\ge \frac{\mu}{\|v\|_1} \ge \frac{\mu}{1+\epsilon}. \end{aligned} \tag{15}$$

Now, define $\bar{\Theta} = \left[\bar{\Delta}^{\epsilon}(\mathscr{S})\right]^{1+(H-1)SA} \times \left[\bar{\Delta}^{\epsilon}(\mathscr{O})\right]^{HS}$. We have

$$\begin{aligned}\log|\bar{\Theta}| &\le [1+(H-1)SA]S\log\left(\frac{S}{\epsilon}+1\right) + HSO\log\left(\frac{O}{\epsilon}+1\right) \\ &\le (HS^2A + HSO)\log\left(\frac{\max\{S,O\}}{\epsilon}+1\right).\end{aligned}$$

For any $\theta \in \Theta$, let $\bar{\theta} \in \bar{\Theta}$ be such that each component of $\bar{\theta}$ is obtained by quantizing and renormalizing the corresponding component in θ according to (14). Then, for any $\tau = (o_h, a_h)_{h=1}^{H} \in \mathscr{T}$, using (2) and (15) we have $\mathbb{P}_{\bar{\theta}}^{-}(\tau) \ge (1+\epsilon)^{-2H}\mathbb{P}_{\theta}^{-}(\tau)$, establishing (11).

Next we establish (10) for θ and $\bar{\theta}$. Define a coupled POMDP with state space $\mathscr{S}^2$, action space $\mathscr{A}^2$, observation space $\mathscr{O}^2$, horizon H as follows

- The initial distribution $\boldsymbol{b}_1^{\theta,\bar{\theta}}$ is an optimal coupling[6] of b_1^{θ} and $b_1^{\bar{\theta}}$;
- The transition kernel $\mathbf{T}^{\theta,\bar{\theta}} : (\mathscr{S}\times\mathscr{A})^2 \mapsto \Delta(\mathscr{S}^2)$ is such that for any $h \in [H-1]$, $s, \bar{s} \in \mathscr{S}$, $a, \bar{a} \in \mathscr{A}$, the joint distribution $\mathbf{T}_h^{\theta,\bar{\theta}}(s,a,\bar{s},\bar{a})$ is an optimal coupling of $T_h^{\theta}(s,a)$ and $T_h^{\bar{\theta}}(\bar{s},\bar{a})$.
- The observation kernel $\mathbf{Z}^{\theta,\bar{\theta}} : \mathscr{S}^2 \mapsto \Delta(\mathscr{O}^2)$ is such that for any $h \in [H]$, $s, \bar{s} \in \mathscr{S}$, $o, \bar{o} \in \mathscr{O}$, the joint distribution $\mathbf{Z}_h^{\theta,\bar{\theta}}(s,\bar{s})$ is an optimal coupling of $Z_h^{\theta}(s)$ and $Z_h^{\bar{\theta}}(\bar{s})$.

In the coupled POMDP, at any given step, if the current states are the same for both components, then the current observation would be the same with probability at least $1-\epsilon$; if the current state and action pairs are the same for both components, then the next states would be the same with probability at least $1-\epsilon$. Consequently, if the same policy $\pi \in \Pi$ is applied on both components, then with probability at least $(1-\epsilon)^{2H} \ge 1-2H\epsilon$, the two trajectories $\tau, \bar{\tau} \in \mathscr{T}$ coincide. Due to the fact that the TV distance equals the minimum mismatch probability of any coupling [Levin and Peres, 2017], we conclude that $\|\mathbb{P}_{\theta}^{\pi} - \mathbb{P}_{\bar{\theta}}^{\pi}\|_{\mathrm{TV}} \le 2H\epsilon$.

D.2 Proof of Lemma D.2 (Lemma 5.1, restated)

First, the episode k regret satisfies

$$\begin{aligned}&\mathbb{E}[V_{\theta^*}^{*}] - \mathbb{E}[V_{\theta^*}^{\tilde{\pi}^k}] \\ &= \mathbb{E}[V_{\tilde{\theta}^k}^{*}] - \mathbb{E}[V_{\theta^*}^{\tilde{\pi}^k}] = \mathbb{E}[V_{\tilde{\theta}^k}^{\tilde{\pi}^k} - V_{\theta^*}^{\tilde{\pi}^k}] \\ &= \mathbb{E}\left[\sum_{\tau=(o_h,a_h)_{h=1}^{H}\in\mathscr{T}} [\mathbb{P}_{\tilde{\theta}^k}^{\tilde{\pi}^k}(\tau) - \mathbb{P}_{\theta^*}^{\tilde{\pi}^k}(\tau)] \sum_{h=1}^{H} r_h(o_h,a_h)\right] \\ &\le H\mathbb{E}\left[\sum_{\tau\in\mathscr{T}} [\mathbb{P}_{\tilde{\theta}^k}^{\tilde{\pi}^k}(\tau) - \mathbb{P}_{\theta^*}^{\tilde{\pi}^k}(\tau)]_+\right] = H\mathbb{E}[\|\mathbb{P}_{\tilde{\theta}^k}^{\tilde{\pi}^k} - \mathbb{P}_{\theta^*}^{\tilde{\pi}^k}\|_{\mathrm{TV}}]\end{aligned}$$

[6] A coupling of two distributions $\nu_1, \nu_2 \in \Delta(\Omega)$ is a joint distribution $\nu_{12} \in \Delta(\Omega\times\Omega)$ with marginal distributions ν_1 and ν_2. An optimal coupling [Levin and Peres, 2017] is a coupling that attains the minimum value of the mismatch probability $\nu_{12}(\{(\omega_1,\omega_2) : \omega_1 \ne \omega_2\})$, which equals $\|\nu_1 - \nu_2\|_{\mathrm{TV}}$. See [Levin and Peres, 2017] for a construction of such coupling.

where in the first identity we used the fact that θ^* has the same unconditional distribution as $\tilde{\theta}^k$. Using Lemma E.3 and the fact that conditioning on $\mathcal{D}_k$, θ^* has the same distribution as $\tilde{\theta}^k$, we have $\ell(\tilde{\theta}^k, \mathcal{D}_k) \geq \bar{\ell}^*(\mathcal{D}_k) - \log(K|\bar{\Theta}|)$ with probability at least $1 - \frac{1}{K}$. Let $\vartheta^k = \iota\left(\tilde{\theta}^k\right)$, we have $\ell(\vartheta^k, \mathcal{D}_k) \geq \ell(\tilde{\theta}^k, \mathcal{D}_k) - 1$ almost surely due to Quick Fact 1. Therefore, we conclude that

$$\mathbb{P}\left(\vartheta^k \in \bar{\Theta}(\mathcal{D}_k)\right) = \mathbb{P}\left(\ell(\vartheta^k, \mathcal{D}_k) \geq \bar{\ell}^*(\mathcal{D}_k) - \log(K|\bar{\Theta}|) - 1\right) \geq 1 - \frac{1}{K}$$

Also note that $\|\mathbb{P}_{\tilde{\theta}^k}^{\tilde{\pi}^k} - \mathbb{P}_{\vartheta^k}^{\tilde{\pi}^k}\|_{\mathrm{TV}} \leq 2H\epsilon = \frac{1}{K}$ almost surely. Therefore

$$\begin{aligned}
\mathbb{E}[\|\mathbb{P}_{\tilde{\theta}^k}^{\tilde{\pi}^k} - \mathbb{P}_{\theta^*}^{\tilde{\pi}^k}\|_{\mathrm{TV}}] &\leq \mathbb{E}[\|\mathbb{P}_{\vartheta^k}^{\tilde{\pi}^k} - \mathbb{P}_{\theta^*}^{\tilde{\pi}^k}\|_{\mathrm{TV}}] + \frac{1}{K} \\
&\leq \mathbb{E}[\|\mathbb{P}_{\vartheta^k}^{\tilde{\pi}^k} - \mathbb{P}_{\theta^*}^{\tilde{\pi}^k}\|_{\mathrm{TV}} \cdot \mathbf{1}_{\{\vartheta^k \in \bar{\Theta}(\mathcal{D}_k)\}}] + \mathbb{P}(\vartheta^k \notin \bar{\Theta}(\mathcal{D}_k)) + \frac{1}{K} \\
&\leq \mathbb{E}\left[\max_{\bar{\theta} \in \bar{\Theta}(\mathcal{D}_k)} \|\mathbb{P}_{\bar{\theta}}^{\tilde{\pi}^k} - \mathbb{P}_{\theta^*}^{\tilde{\pi}^k}\|_{\mathrm{TV}}\right] + \frac{2}{K}
\end{aligned}$$

and finally

$$\begin{aligned}
\mathrm{BReg}(\phi^{\mathrm{PS4POMDP}}, K) &\leq \sum_{k=1}^{K} H\left(\mathbb{E}\left[\max_{\bar{\theta} \in \bar{\Theta}(\mathcal{D}_k)} \|\mathbb{P}_{\bar{\theta}}^{\tilde{\pi}^k} - \mathbb{P}_{\theta^*}^{\tilde{\pi}^k}\|_{\mathrm{TV}}\right] + \frac{2}{K}\right) \\
&\leq 2H + H\mathbb{E}\left[\sum_{k=1}^{K} \max_{\bar{\theta} \in \bar{\Theta}(\mathcal{D}_k)} \|\mathbb{P}_{\bar{\theta}}^{\tilde{\pi}^k} - \mathbb{P}_{\theta^*}^{\tilde{\pi}^k}\|_{\mathrm{TV}}\right].
\end{aligned}$$

D.3 Proof of Lemma D.3 (Lemma 5.2, restated)

Consider the distribution on the learning process conditioning on $\theta^* = \theta$. Fix $\theta' \in \bar{\Theta}$. Define $\mathcal{F}_{k-1}$ to be the σ-algebra generated from $(\mathcal{D}_k, \pi^k)$ and $Z_k = \log\left(\frac{\mathbb{P}_{\theta'}^{\pi^k}(\tau^k)}{\mathbb{P}_{\theta}^{\pi^k}(\tau^k)}\right)$. Then the quantity ψ_k defined in Lemma E.1 satisfies

$$\begin{aligned}
\psi_k\left(\frac{1}{2}\right) &= \log \mathbb{E}_\theta\left[\exp\left(\frac{1}{2}\log\left(\frac{\mathbb{P}_{\theta'}^{\pi^k}(\tau^k)}{\mathbb{P}_{\theta}^{\pi^k}(\tau^k)}\right)\right) \;\middle|\; \mathcal{D}_k, \pi^k\right] \\
&= \log\left(\sum_{\tau \in \mathscr{T}: \mathbb{P}_{\theta}^{\pi^k}(\tau) > 0} \mathbb{P}_{\theta}^{\pi^k}(\tau)\sqrt{\frac{\mathbb{P}_{\theta'}^{\pi^k}(\tau)}{\mathbb{P}_{\theta}^{\pi^k}(\tau)}}\right) = \log\left(\sum_{\tau \in \mathscr{T}} \sqrt{\mathbb{P}_{\theta'}^{\pi^k}(\tau)\mathbb{P}_{\theta}^{\pi^k}(\tau)}\right) \\
&\leq \left(\sum_{\tau \in \mathscr{T}} \sqrt{\mathbb{P}_{\theta'}^{\pi^k}(\tau)\mathbb{P}_{\theta}^{\pi^k}(\tau)}\right) - 1 = \sum_{\tau \in \mathscr{T}}\left(\sqrt{\mathbb{P}_{\theta'}^{\pi^k}(\tau)\mathbb{P}_{\theta}^{\pi^k}(\tau)} - \frac{1}{2}\mathbb{P}_{\theta'}^{\pi^k}(\tau) - \frac{1}{2}\mathbb{P}_{\theta}^{\pi^k}(\tau)\right) \\
&= -\frac{1}{2}\sum_{\tau \in \mathscr{T}}\left(\sqrt{\mathbb{P}_{\theta'}^{\pi^k}(\tau)} - \sqrt{\mathbb{P}_{\theta}^{\pi^k}(\tau)}\right)^2 \leq -\frac{1}{2}\|\mathbb{P}_{\theta'}^{\pi^k} - \mathbb{P}_{\theta}^{\pi^k}\|_{\mathrm{TV}}^2
\end{aligned}$$

where we have used the inequality $\log(t) \leq t - 1$, and the last inequality is due to Lemma E.2. Applying Lemma E.1, for any $\xi > 0$, with probability at least $1 - e^{-\xi}$,

$$\sum_{j=1}^{k-1}\left(\frac{1}{2}Z_j + \frac{1}{2}\|\mathbb{P}_{\theta'}^{\pi^j} - \mathbb{P}_{\theta}^{\pi^j}\|_{\mathrm{TV}}^2\right) \leq \sum_{j=1}^{k-1}\left(\frac{1}{2}Z_j - \psi_j\left(\frac{1}{2}\right)\right) < \xi \qquad \forall k \in [K]$$

or equivalently,

$$\sum_{j=1}^{k-1}\|\mathbb{P}_{\theta'}^{\pi^j} - \mathbb{P}_{\theta}^{\pi^j}\|_{\mathrm{TV}}^2 < 2\xi - \sum_{j=1}^{k-1} Z_j = 2\xi + \ell(\theta, \mathcal{D}_k) - \ell(\theta', \mathcal{D}_k) \qquad \forall k \in [K] \tag{16}$$

Using the fact that TV distances are bounded by 1, adding in the k-th term, (16) implies

$$\sum_{j=1}^{k} \|\mathbb{P}_{\theta'}^{\pi^j} - \mathbb{P}_{\theta}^{\pi^j}\|_{\mathrm{TV}}^2 < 2\xi + 1 + \ell(\theta, \mathcal{D}_k) - \ell(\theta', \mathcal{D}_k) \qquad \forall k \in [K]$$

Now, if $\theta' \in \bar{\Theta}(\mathcal{D}_k)$, then

$$\ell(\theta, \mathcal{D}_k) - \ell(\theta', \mathcal{D}_k) \le \bar{\ell}^*(\mathcal{D}_k) + 1 - \ell(\theta', \mathcal{D}_k) \le \log(K|\bar{\Theta}|) + 2$$

where the first inequality is due to Quick Fact 2, and the second inequality is due to the definition of $\bar{\Theta}(\mathcal{D}_k)$.

Finally, we have

$$\begin{aligned}
&\mathbb{P}_\theta\left(\max_{k\in[K]}\max_{\theta'\in\bar{\Theta}(\mathcal{D}_k)}\sum_{j=1}^{k}\|\mathbb{P}_{\theta'}^{\pi^j} - \mathbb{P}_{\theta}^{\pi^j}\|_{\mathrm{TV}}^2 \ge 3\log\left(K|\bar{\Theta}|\right) + 3\right)\\
&= \mathbb{P}_\theta\left(\exists k\in[K], \theta'\in\bar{\Theta}(\mathcal{D}_k),\ \sum_{j=1}^{k}\|\mathbb{P}_{\theta'}^{\pi^j} - \mathbb{P}_{\theta}^{\pi^j}\|_{\mathrm{TV}}^2 \ge 3\log\left(K|\bar{\Theta}|\right) + 3\right)\\
&\le \mathbb{P}_\theta\left(\exists k\in[K], \theta'\in\bar{\Theta},\ \sum_{j=1}^{k}\|\mathbb{P}_{\theta'}^{\pi^j} - \mathbb{P}_{\theta}^{\pi^j}\|_{\mathrm{TV}}^2 \ge 2\log\left(K|\bar{\Theta}|\right) + 1 + \ell(\theta, \mathcal{D}_k) - \ell(\theta', \mathcal{D}_k)\right)\\
&\le \sum_{\theta'\in\bar{\Theta}}\mathbb{P}_\theta\left(\exists k\in[K],\ \sum_{j=1}^{k}\|\mathbb{P}_{\theta'}^{\pi^j} - \mathbb{P}_{\theta}^{\pi^j}\|_{\mathrm{TV}}^2 \ge 2\log\left(K|\bar{\Theta}|\right) + 1 + \ell(\theta, \mathcal{D}_k) - \ell(\theta', \mathcal{D}_k)\right)\\
&\le \sum_{\theta'\in\bar{\Theta}}\exp\left(-\log(K|\bar{\Theta}|)\right) = \frac{1}{K}.
\end{aligned}$$

The above holds for any $\theta \in \Theta$. Therefore, we conclude that with probability at least $1 - \frac{1}{K}$,

$$\max_{k\in[K]}\max_{\theta'\in\bar{\Theta}(\mathcal{D}_k)}\sum_{j=1}^{k}\|\mathbb{P}_{\theta'}^{\pi^j} - \mathbb{P}_{\theta^*}^{\pi^j}\|_{\mathrm{TV}}^2 < 3\log\left(K|\bar{\Theta}|\right) + 3.$$

D.4 Proof of Theorem 4.1

Define $(OA)^H$-dimensional vectors x_k, w_k via

$$\begin{aligned}
x_k(\tau) &:= \tilde{\pi}^k(\tau)\\
w_k(\tau) &:= \frac{1}{2}|\mathbb{P}_{\tilde{\theta}^k}^{-}(\tau) - \mathbb{P}_{\theta^*}^{-}(\tau)|
\end{aligned}$$

for each trajectory $\tau \in \mathscr{T}$, then (TR-MLE) can be written as $\sum_{j=1}^{k}(w_k^T x_j)^2$ and (TR-REG) can be written as $\sum_{k=1}^{K} w_k^T x_k$. We have

$$\begin{aligned}
\|x_k\|_2 \le \|x_k\|_1 &= \sum_{\tau\in\mathscr{T}}\tilde{\pi}^k(\tau) = O^H\\
\|w_k\|_2 \le \|w_k\|_1 &= \sum_{\tau\in\mathscr{T}}\frac{1}{2}|\mathbb{P}_{\tilde{\theta}^k}^{-}(\tau) - \mathbb{P}_{\theta^*}^{-}(\tau)| = A^H\sum_{\tau\in\mathscr{T}}\frac{1}{2}|\mathbb{P}_{\tilde{\theta}^k}^{\mathbf{unif}}(\tau) - \mathbb{P}_{\theta^*}^{\mathbf{unif}}(\tau)| \le A^H
\end{aligned}$$

where $\mathbf{unif}$ represents the policy of picking a uniform random action at all times.

Lemma D.3, in particular, implies that $\sum_{j=1}^{k}(w_k^T x_j)^2 \le 3\log(K|\bar{\Theta}|) + 3$ with probability at least $1 - \frac{1}{K}$.

Applying Lemma 5.3 with $\lambda = 1$, with probability $1 - \frac{1}{K}$, we obtain

$$\text{(TR-REG)} \le \sqrt{(1+\beta)(OA)^H K\log\left[1 + (OA)^H K\right]}$$

where $\beta := 3\log(K|\bar{\Theta}|) + 3 = \tilde{\mathcal{O}}(HS^2A + HSO)$.

Since TV distances are bounded by 1, we also have (TR-REG) $\leq K$ with probability 1, hence

$$\mathbb{E}[\text{(TR-REG)}] \leq \sqrt{(1+\beta)(OA)^H K \log\left[1 + (OA)^H K\right]} + 1$$

Finally, using Lemma D.2, we conclude that

$$\begin{aligned}
\mathrm{BReg}(\phi^{\mathrm{PS4POMDP}}, K) &\leq 2H + H\mathbb{E}\left[\text{(TR-REG)}\right] \\
&\leq 2H + H\sqrt{(1+\beta)(OA)^H K \log\left[1 + (OA)^H K\right]} + H \\
&= \tilde{\mathcal{O}}\left(H\sqrt{H\beta(OA)^H K}\right) = \tilde{\mathcal{O}}\left(H^2\sqrt{(S^2A + SO)(OA)^H K}\right).
\end{aligned}$$

D.5 Proof of Theorem 4.4

To obtain a bound polynomial in H, we will follow the indirect approach as illustrated in Figure 2 and incorporate Assumption 3.1 along the way. The steps here follow closely that of [Liu et al., 2022]. However, through the use of an index change lemma with a tighter upper bound guarantee (our Proposition B.2 in Appendix B, compared to Proposition 22 by [Liu et al., 2022]), we are able to establish a regret bound for weakly α-revealing POMDPs with improved dependency on H, S, and A.

Before we start the proof, we first establish a few quick results from Assumption 3.1:

Quick Fact 3: $\alpha \leq \sqrt{S}$.

Proof. We have $\sigma_{\min}(\mathbb{Z}_h^\theta) \leq \|\mathbb{Z}_h^\theta\|_2 \leq \sqrt{S}\|\mathbb{Z}_h^\theta\|_1$. Furthermore, $\|\mathbb{Z}_h^\theta\|_1 = \max_{s\in\mathscr{S}} \sum_{o\in\mathscr{O}} |\mathbb{Z}_h^\theta(o,s)| = 1$. □

Quick Fact 4: Let $\mathbb{Z}_h^{\theta\dagger}$ be the Moore-Penrose inverse of $\mathbb{Z}_h^\theta$. We have $\|\mathbb{Z}_h^{\theta\dagger}\|_1 \leq \dfrac{\sqrt{S}}{\alpha}$ for $\theta \in \Theta$.

Proof. $\|\mathbb{Z}_h^{\theta\dagger}\|_1 \leq \sqrt{S}\|\mathbb{Z}_h^{\theta\dagger}\|_2 = \sqrt{S}\sigma_{\max}(\mathbb{Z}_h^{\theta\dagger}) = \dfrac{\sqrt{S}}{\sigma_{\min}(\mathbb{Z}_h^\theta)} \leq \dfrac{\sqrt{S}}{\alpha}$. □

Similarly to the proof of Theorem 4.1, in this proof, we attempt to bound (TR-REG) with (TR-MLE). Note that while Assumption 3.1 is imposed on the set Θ, both (TR-REG) and (TR-MLE) involve some $\bar{\theta} \in \bar{\Theta}$. The fact that $\bar{\Theta}$ is not a subset of Θ prevents us from applying Assumption 3.1. Therefore, we need the following step. Recall that we have assumed that the quantization map ι is surjective. Let $\hat{\theta}^k \in \Theta$ be such that $\check{\theta}^k = \iota(\hat{\theta}^k)$, then we have $\|\mathbb{P}^\pi_{\check{\theta}^k} - \mathbb{P}^\pi_{\hat{\theta}^k}\|_{\mathrm{TV}} \leq 2H\epsilon = \frac{1}{K}$. Therefore, we have

$$\begin{aligned}
\sum_{j=1}^k \|\mathbb{P}^{\tilde{\pi}^j}_{\hat{\theta}^k} - \mathbb{P}^{\tilde{\pi}^j}_{\theta^*}\|^2_{\mathrm{TV}} &\leq \sum_{j=1}^k \left(\|\mathbb{P}^{\tilde{\pi}^j}_{\check{\theta}^k} - \mathbb{P}^{\tilde{\pi}^j}_{\theta^*}\|_{\mathrm{TV}} + \frac{1}{K}\right)^2 = \sum_{j=1}^k \left(\|\mathbb{P}^{\tilde{\pi}^j}_{\check{\theta}^k} - \mathbb{P}^{\tilde{\pi}^j}_{\theta^*}\|^2_{\mathrm{TV}} + \frac{2}{K}\|\mathbb{P}^{\tilde{\pi}^j}_{\check{\theta}^k} - \mathbb{P}^{\tilde{\pi}^j}_{\theta^*}\|_{\mathrm{TV}} + \frac{1}{K^2}\right) \\
&\leq \left(\sum_{j=1}^k \|\mathbb{P}^{\tilde{\pi}^j}_{\check{\theta}^k} - \mathbb{P}^{\tilde{\pi}^j}_{\theta^*}\|^2_{\mathrm{TV}}\right) + k\left(\frac{2}{K} + \frac{1}{K^2}\right) \leq \text{(TR-MLE)} + 2 + \frac{1}{K} \qquad (17)
\end{aligned}$$

and

$$\text{(TR-REG)} \leq \sum_{j=1}^K \left(\|\mathbb{P}^{\tilde{\pi}^j}_{\hat{\theta}^j} - \mathbb{P}^{\tilde{\pi}^j}_{\theta^*}\|_{\mathrm{TV}} + \frac{1}{K}\right) \leq \left(\sum_{j=1}^K \|\mathbb{P}^{\tilde{\pi}^j}_{\hat{\theta}^j} - \mathbb{P}^{\tilde{\pi}^j}_{\theta^*}\|_{\mathrm{TV}}\right) + 1. \qquad (18)$$

Therefore, our task for this proof is to use

$$\sum_{j=1}^{k} \|\mathbb{P}_{\hat{\theta}^k}^{\tilde{\pi}^j} - \mathbb{P}_{\theta^*}^{\tilde{\pi}^j}\|_{\mathrm{TV}}^2 \tag{TR-MLE-II}$$

to bound

$$\sum_{j=1}^{K} \|\mathbb{P}_{\hat{\theta}^j}^{\tilde{\pi}^j} - \mathbb{P}_{\theta^*}^{\tilde{\pi}^j}\|_{\mathrm{TV}}. \tag{TR-REG-II}$$

Step 1: Bound (TR-REG-II) **with Projected Distances of Observable Operators.**

Recall that the probability of a trajectory of a POMDP can be written with matrix multiplications. Recall that $\mathbb{T}_{h,a} \in \mathbb{R}^{\mathscr{S}\times\mathscr{S}}$ is the probability transition matrix (where the rows represent the next state) under action $a \in \mathscr{A}$ at time $h \in [H-1]$. Recall that $\mathbb{Z}_h \in \mathbb{R}^{\mathscr{O}\times\mathscr{S}}$ is the observation probability matrix. For each $h \in [H-1], a \in \mathscr{A}, o \in \mathscr{O}$, define the *observable operator* matrix [Jin et al., 2020]

$$\mathbf{B}_h^\theta(a,o) := \mathbb{Z}_{h+1}^\theta \mathbb{T}_{h,a}^\theta \mathrm{diag}(\mathbb{Z}_h^\theta(o,\cdot))\mathbb{Z}_h^{\theta\dagger} \tag{19}$$

where $\mathbb{Z}_h^{\theta\dagger}$ is the Moore-Penrose inverse of $\mathbb{Z}_h^\theta$. Note that due to Assumption 3.1, $\mathbb{Z}_h^\theta$ has full column rank, hence $\mathbb{Z}_h^{\theta\dagger}\mathbb{Z}_h^\theta = \mathbf{I}$. We can rewrite the environment part of the probability of a trajectory as

$$\mathbb{P}_\theta^-(o_{1:H}, a_{1:H}) = \mathbf{e}_{o_H}^T \mathbf{B}_{H-1}^\theta(a_{H-1}, o_{H-1}) \cdots \mathbf{B}_1^\theta(a_1, o_1)\mathbb{Z}_1^\theta b_1^\theta \tag{20}$$

For a full trajectory $\tau = (o_{1:H}, a_{1:H}) \in \mathscr{T}$, let $\tau_h = (o_{1:h}, a_{1:h})$ denote its corresponding partial trajectory up to time h and $\tau_{-h} = (o_{h+1:H}, a_{h+1:H})$ denote its corresponding partial trajectory from time $h+1$ to H. Define $\mathbf{v}_0^\theta = \mathbb{Z}_1^\theta b_1^\theta$ and $\mathbf{v}_h^\theta(\tau_h) := \mathbf{B}_h^\theta(a_h, o_h) \cdots \mathbf{B}_1^\theta(a_1, o_1)\mathbf{v}_0^\theta$. With some abuse of notation, define

$$\mathbf{B}_{H:h}^\theta(\tau_{-(h-1)}) := \begin{cases} \mathbf{e}_{o_H}^T & h = H \\ \mathbf{e}_{o_H}^T \mathbf{B}_{H-1}^\theta(a_{H-1}, o_{H-1}) \cdots \mathbf{B}_h^\theta(a_h, o_h) & h < H \end{cases} \tag{21}$$

In the rest of the step, we will prove the following claim.

Claim: With probability 1, we have (TR-REG-II) $\leq \frac{\sqrt{S}}{\alpha} \sum_{h=0}^{H-1}$(OP-REG-$h$), where

$$\text{(OP-REG-0)} := \frac{1}{2}\sum_{j=1}^{K} \|\mathbf{v}_0^{\theta^*} - \mathbf{v}_0^{\hat{\theta}^j}\|_1$$

$$\text{(OP-REG-}h\text{)} := \frac{1}{2}\sum_{j=1}^{K} \sum_{\tau_h \in (\mathscr{O}\times\mathscr{A})^h} \tilde{\pi}^j(\tau_h) \|[\mathbf{B}_h^{\theta^*}(a_h, o_h) - \mathbf{B}_h^{\hat{\theta}^j}(a_h, o_h)]\mathbf{v}_{h-1}^{\theta^*}(\tau_{h-1})\|_1 \qquad \forall h \in [H-1]$$

are the *projected operator distances*.

From (20), using telescoping sums, for any $\pi \in \mathit{\Pi}$ and parameters θ, θ' we have

$$\begin{aligned} \|\mathbb{P}_\theta^\pi - \mathbb{P}_{\theta'}^\pi\|_{\mathrm{TV}} &= \frac{1}{2}\sum_{\tau\in\mathscr{T}} \pi(\tau)|\mathbb{P}_\theta^-(\tau) - \mathbb{P}_{\theta'}^-(\tau)| \\ \leq \frac{1}{2}\sum_{\tau\in\mathscr{T}} \pi(\tau)|\mathbf{B}_{H:1}^{\theta'}(\tau)[\mathbf{v}_0^\theta - \mathbf{v}_0^{\theta'}]| \\ + \frac{1}{2}\sum_{\tau\in\mathscr{T}} \pi(\tau) \sum_{h=1}^{H-1} |\mathbf{B}_{H:h+1}^{\theta'}(\tau_{-h})[\mathbf{B}_h^\theta(a_h, o_h) - \mathbf{B}_h^{\theta'}(a_h, o_h)]\mathbf{v}_{h-1}^\theta(\tau_{h-1})| \end{aligned}$$

Using Lemma E.5, we have

$$\sum_{\tau\in\mathscr{T}} \pi(\tau)|\mathbf{B}_{H:1}^{\theta'}(\tau)[\mathbf{v}_0^\theta - \mathbf{v}_0^{\theta'}]| \leq \frac{\sqrt{S}}{\alpha}\|\mathbf{v}_0^\theta - \mathbf{v}_0^{\theta'}\|_1$$

Also through Lemma E.5, for each $h \in [H-1]$ we have

$$\begin{aligned}
&\sum_{\tau \in \mathscr{T}} \pi(\tau) |\mathbf{B}^{\theta'}_{H:h+1}(\tau_{-h})[\mathbf{B}^{\theta}_h(a_h, o_h) - \mathbf{B}^{\theta'}_h(a_h, o_h)]\mathbf{v}^{\theta}_{h-1}(\tau_{h-1})| \\
= &\sum_{\tau_h \in (\mathscr{O}\times\mathscr{A})^h : \pi(\tau_h) > 0} \pi(\tau_h) \sum_{\tau_{-h} \in (\mathscr{O}\times\mathscr{A})^{H-h}} \pi(\tau_{-h}|\tau_h) |\mathbf{B}^{\theta'}_{H:h+1}(\tau_{-h})[\mathbf{B}^{\theta}_h(a_h, o_h) - \mathbf{B}^{\theta'}_h(a_h, o_h)]\mathbf{v}^{\theta}_{h-1}(\tau_{h-1})| \\
\leq &\sum_{\tau_h \in (\mathscr{O}\times\mathscr{A})^h : \pi(\tau_h) > 0} \pi(\tau_h) \cdot \frac{\sqrt{S}}{\alpha} \|[\mathbf{B}^{\theta}_h(a_h, o_h) - \mathbf{B}^{\theta'}_h(a_h, o_h)]\mathbf{v}^{\theta}_{h-1}(\tau_{h-1})\|_1
\end{aligned}$$

where for each $\pi \in \Pi$ and $\tau_h \in (\mathscr{O} \times \mathscr{A})^h$, $\pi(\tau_h)$ is defined in a similar way as $\pi(\tau)$ except that τ_h is a partial trajectory. For π, τ_h such that $\pi(\tau_h) > 0$, $\pi(\tau_{-h}|\tau_h)$ is defined as $\pi(\tau_h, \tau_{-h})/\pi(\tau_h)$.

We conclude that (TR-REG-II) can be bounded by

$$\begin{aligned}
&\sum_{j=1}^{K} \|\mathbb{P}^{\tilde{\pi}^j}_{\hat{\theta}^j} - \mathbb{P}^{\tilde{\pi}^j}_{\theta^*}\|_{\mathrm{TV}} \\
\leq &\frac{\sqrt{S}}{2\alpha} \sum_{j=1}^{K} \left(\|\mathbf{v}^{\theta^*}_0 - \mathbf{v}^{\hat{\theta}^j}_0\|_1 + \sum_{h=1}^{H-1} \sum_{\tau_h \in (\mathscr{O}\times\mathscr{A})^h} \tilde{\pi}^j(\tau_h) \left\| [\mathbf{B}^{\theta^*}_h(a_h, o_h) - \mathbf{B}^{\hat{\theta}^j}_h(a_h, o_h)]\mathbf{v}^{\theta^*}_{h-1}(\tau_{h-1}) \right\|_1 \right)
\end{aligned}$$

proving the claim.

Step 2: Bound Projected Distances of Observable Operators with (TR-MLE-II)**.** In this step we prove the following claim

Claim: With probability at least $1 - \frac{1}{K}$,

$$\text{(OP-MLE-0)} := \frac{1}{2}\|\mathbf{v}^{\hat{\theta}^k}_0 - \mathbf{v}^{\theta^*}_0\|_1 \leq \sqrt{\frac{\beta}{k}} \tag{22}$$

$$\text{(OP-MLE-}h\text{)} := \sum_{j=1}^{k} \left(\frac{1}{2} \sum_{\tau_h} \tilde{\pi}^j(\tau_h) \|[\mathbf{B}^{\theta^*}_h(a_h, o_h) - \mathbf{B}^{\hat{\theta}^k}_h(a_h, o_h)]\mathbf{v}^{\theta^*}_{h-1}(\tau_{h-1})\|_1 \right)^2 \leq \frac{4S}{\alpha^2}\beta \tag{23}$$

for all $k \in [K]$, where $\beta := 3\log(K|\bar{\Theta}|) + 5 + \dfrac{1}{K}$.

Observe that the vectors $\mathbf{v}^{\theta}_0 \in \mathbb{R}^{\mathscr{O}}$ and $\mathbf{v}^{\theta}_h(\tau_h) \in \mathbb{R}^{\mathscr{O}}$ defined in Step 1 have the following probabilistic interpretation

$$\begin{aligned}
\mathbf{v}^{\theta}_0(o_1) = \mathbb{P}_{\theta}(o_1) \qquad &\forall o_1 \in \mathscr{O} \\
\pi(\tau_h) \cdot [\mathbf{v}^{\theta}_h(\tau_h)](o_{h+1}) = \mathbb{P}^{\pi}_{\theta}(\tau_h, o_{h+1}) \qquad &\forall \tau_h \in (\mathscr{O} \times \mathscr{A})^h, o_{h+1} \in \mathscr{O}
\end{aligned}$$

Therefore, for all $\theta, \theta' \in \Theta, \pi \in \Pi$, we have

$$\begin{aligned}
\frac{1}{2}\|\mathbf{v}^{\theta}_0 - \mathbf{v}^{\theta'}_0\|_1 = \frac{1}{2}\sum_{o_1} |\mathbb{P}_{\theta}(o_1) - \mathbb{P}_{\theta'}(o_1)| \leq \|\mathbb{P}^{\pi}_{\theta} - \mathbb{P}^{\pi}_{\theta'}\|_{\mathrm{TV}} \\
\frac{1}{2}\sum_{\tau_h} \pi(\tau_h)\|\mathbf{v}^{\theta}_h(\tau_h) - \mathbf{v}^{\theta'}_h(\tau_h)\|_1 = \frac{1}{2}\sum_{\tau_h}\sum_{o_{h+1}} |\mathbb{P}^{\pi}_{\theta}(\tau_h, o_{h+1}) - \mathbb{P}^{\pi}_{\theta'}(\tau_h, o_{h+1})| \leq \|\mathbb{P}^{\pi}_{\theta} - \mathbb{P}^{\pi}_{\theta'}\|_{\mathrm{TV}}
\end{aligned}$$

where we have used the fact that the total variation distance between two marginal distributions is no more than that between the original distributions. Therefore, we have

$$k\left(\frac{1}{2}\|\mathbf{v}^{\hat{\theta}^k}_0 - \mathbf{v}^{\theta^*}_0\|_1\right)^2 = \sum_{j=1}^{k}\left(\frac{1}{2}\|\mathbf{v}^{\hat{\theta}^k}_0 - \mathbf{v}^{\theta^*}_0\|_1\right)^2 \leq \text{(TR-MLE-II)} \tag{24}$$

$$\sum_{j=1}^{k}\left(\frac{1}{2}\sum_{\tau_h}\tilde{\pi}^j(\tau_h)\|\mathbf{v}_h^{\hat{\theta}^k}(\tau_h)-\mathbf{v}_h^{\theta^*}(\tau_h)\|_1\right)^2 \leq \text{(TR-MLE-II)} \tag{25}$$

We proceed to bound the projected operator distance. First, for any $\tau_h \in (\mathscr{O}\times\mathscr{A})^h$ we have

$$\begin{aligned}
&\|[\mathbf{B}_h^{\theta^*}(a_h,o_h)-\mathbf{B}_h^{\hat{\theta}^k}(a_h,o_h)]\mathbf{v}_{h-1}^{\theta^*}(\tau_{h-1})\|_1 \\
&\leq \|\mathbf{B}_h^{\theta^*}(a_h,o_h)\mathbf{v}_{h-1}^{\theta^*}(\tau_{h-1})-\mathbf{B}_h^{\hat{\theta}^k}(a_h,o_h)\mathbf{v}_{h-1}^{\hat{\theta}^k}(\tau_{h-1})\|_1+\|\mathbf{B}_h^{\hat{\theta}^k}(a_h,o_h)[\mathbf{v}_{h-1}^{\hat{\theta}^k}(\tau_{h-1})-\mathbf{v}_{h-1}^{\theta^*}(\tau_{h-1})]\|_1 \\
&= \|\mathbf{v}_h^{\theta^*}(\tau_h)-\mathbf{v}_h^{\hat{\theta}^k}(\tau_h)\|_1+\|\mathbf{B}_h^{\hat{\theta}^k}(a_h,o_h)[\mathbf{v}_{h-1}^{\hat{\theta}^k}(\tau_{h-1})-\mathbf{v}_{h-1}^{\theta^*}(\tau_{h-1})]\|_1
\end{aligned}$$

where $\tau_{h-1}\in(\mathscr{O}\times\mathscr{A})^{h-1}$ is the partial trajectory made of the first $h-1$ observations and actions in τ_h.

Therefore, using the inequality $(x_1+x_2)^2\leq 2x_1^2+2x_2^2$, we have

$$\begin{aligned}
&\sum_{j=1}^{k}\left(\frac{1}{2}\sum_{\tau_h}\tilde{\pi}^j(\tau_h)\|[\mathbf{B}_h^{\theta^*}(a_h,o_h)-\mathbf{B}_h^{\hat{\theta}^k}(a_h,o_h)]\mathbf{v}_{h-1}^{\theta^*}(\tau_{h-1})\|_1\right)^2 \\
&\leq 2\underbrace{\sum_{j=1}^{k}\left(\frac{1}{2}\sum_{\tau_h}\tilde{\pi}^j(\tau_h)\|\mathbf{v}_h^{\theta^*}(\tau_h)-\mathbf{v}_h^{\hat{\theta}^k}(\tau_h)\|_1\right)^2}_{=:\xi_1} \\
&+2\underbrace{\sum_{j=1}^{k}\left(\frac{1}{2}\sum_{\tau_h}\tilde{\pi}^j(\tau_h)\|\mathbf{B}_h^{\hat{\theta}^k}(a_h,o_h)[\mathbf{v}_{h-1}^{\hat{\theta}^k}(\tau_{h-1})-\mathbf{v}_{h-1}^{\theta^*}(\tau_{h-1})]\|_1\right)^2}_{\xi_2}
\end{aligned}$$

From (25) we know that $\xi_1\leq$ (TR-MLE-II). It remains to bound ξ_2. We have

$$\begin{aligned}
&\frac{1}{2}\sum_{\tau_h}\tilde{\pi}^j(\tau_h)\|\mathbf{B}_h^{\hat{\theta}^k}(a_h,o_h)[\mathbf{v}_{h-1}^{\hat{\theta}^k}(\tau_{h-1})-\mathbf{v}_{h-1}^{\theta^*}(\tau_{h-1})]\|_1 \\
&=\frac{1}{2}\sum_{\tau_{h-1}}\tilde{\pi}^j(\tau_{h-1})\sum_{a_h,o_h}\tilde{\pi}_h^j(a_h|\tau_{h-1},o_h)\|\mathbf{B}_h^{\hat{\theta}^k}(a_h,o_h)[\mathbf{v}_{h-1}^{\hat{\theta}^k}(\tau_{h-1})-\mathbf{v}_{h-1}^{\theta^*}(\tau_{h-1})]\|_1 \\
&\overset{(*)}{\leq}\frac{\sqrt{S}}{2\alpha}\sum_{\tau_{h-1}}\tilde{\pi}^j(\tau_{h-1})\|\mathbf{v}_{h-1}^{\hat{\theta}^k}(\tau_{h-1})-\mathbf{v}_{h-1}^{\theta^*}(\tau_{h-1})\|_1
\end{aligned}$$

where in (*) we have used Lemma E.4.

Therefore, using (25), we have

$$\xi_2\leq\sum_{j=1}^{k}\left(\frac{\sqrt{S}}{2\alpha}\sum_{\tau_{h-1}}\tilde{\pi}^j(\tau_{h-1})\|\mathbf{v}_{h-1}^{\hat{\theta}^k}(\tau_{h-1})-\mathbf{v}_{h-1}^{\theta^*}(\tau_{h-1})\|_1\right)^2\leq\frac{S}{\alpha^2}\text{(TR-MLE-II)}$$

and consequently

$$\begin{aligned}
&\sum_{j=1}^{k}\left(\frac{1}{2}\sum_{\tau_h}\tilde{\pi}^j(\tau_h)\|[\mathbf{B}_h^{\theta^*}(a_h,o_h)-\mathbf{B}_h^{\hat{\theta}^k}(a_h,o_h)]\mathbf{v}_{h-1}^{\theta^*}(\tau_{h-1})\|_1\right)^2 \\
&\leq 2(\xi_1+\xi_2)\leq 2\left(1+\frac{S}{\alpha^2}\right)\text{(TR-MLE-II)}\leq\frac{4S}{\alpha^2}\text{(TR-MLE-II)}.
\end{aligned} \tag{26}$$

From Lemma D.3 and (17), we know that with probability at least $1-\frac{1}{K}$, we have (TR-MLE-II) $\leq 3\log(K|\bar{\Theta}|)+5+\frac{1}{K}=\beta$. The claim then follows from (24) and (26).

Step 3: Bound (OP-REG) with Results on (OP-MLE) via Index Change Lemma. In the final step, for each $h = 0, 1, \cdots, H-1$, we utilize Proposition B.2 to provide an upper bound of each (OP-REG-h) via the high-probability bound on (OP-MLE-h).

Suppose that both (22) and (23) are true. First, we have

$$\text{(OP-REG-0)} = \sum_{k=1}^{K} \frac{1}{2} \|\mathbf{v}_0^{\hat{\theta}^k} - \mathbf{v}_0^{\theta^*}\|_1 \le \sum_{k=1}^{K} \sqrt{\frac{\beta}{k}} \le 2\sqrt{\beta K}.$$

Now, fix $h \in [H-1]$, we will use (OP-MLE-h) to bound (OP-REG-h).

Define the following quantities: For $k \in [K], a \in \mathscr{A}, o, \tilde{o} \in \mathscr{O}, \tau_{h-1} \in (\mathscr{O} \times \mathscr{A})^{h-1}$, define $w_{k,a,o,\tilde{o}}, x_{k,a,o,\tau_{h-1}} \in \mathbb{R}^{\mathscr{S}}$ by

$$w_{k,a,o,\tilde{o}}^T := \frac{1}{2} \mathbf{e}_{\tilde{o}}^T [\mathbf{B}_h^{\theta^*}(a,o) - \mathbf{B}_h^{\hat{\theta}^k}(a,o)] \mathbb{Z}_h^{\theta^*}$$
$$x_{k,a,o,\tau_{h-1}} := \tilde{\pi}^k(\tau_h) \mathbb{Z}_h^{\theta^*\dagger} \mathbf{v}_{h-1}^{\theta^*}(\tau_{h-1})$$

where $\tau_h \in (\mathscr{O} \times \mathscr{A})^h$ is a trajectory such that $a_h = a, o_h = o$, and the first $h-1$ steps are made of τ_{h-1}. Then, (OP-MLE-h) can be written as

$$\sum_{j=1}^{k} \left(\sum_{(a,o) \in \mathscr{A} \times \mathscr{O}} \sum_{\tilde{o} \in \mathscr{O}} \sum_{\tau_{h-1} \in (\mathscr{O} \times \mathscr{A})^{h-1}} |w_{k,a,o,\tilde{o}}^T x_{j,a,o,\tau_{h-1}}| \right)^2.$$

Note that we have used the fact that $\mathbf{v}_{h-1}^{\theta^*}(\tau_{h-1})$ is in the (column) span of $\mathbb{Z}_h^{\theta^*}$ (and hence $\mathbf{v}_{h-1}^{\theta^*}(\tau_{h-1}) = \mathbb{Z}_h^{\theta^*} \mathbb{Z}_h^{\theta^*\dagger} \mathbf{v}_{h-1}^{\theta^*}(\tau_{h-1})$).

Furthermore, for each $k \in [K]$, we have

$$\begin{aligned}
&\sum_{(a,o) \in \mathscr{A} \times \mathscr{O}} \sum_{\tau_{h-1} \in (\mathscr{O} \times \mathscr{A})^{h-1}} \|x_{k,a,o,\tau_{h-1}}\|_1 \\
&= \sum_{(a,o) \in \mathscr{A} \times \mathscr{O}} \sum_{\tau_{h-1} \in (\mathscr{O} \times \mathscr{A})^{h-1}} \|\tilde{\pi}^k(\tau_h) \mathbb{Z}_h^{\theta^*\dagger} \mathbf{v}_{h-1}^{\theta^*}(\tau_{h-1})\|_1 \\
&= \sum_{(a,o) \in \mathscr{A} \times \mathscr{O}} \sum_{\tau_{h-1} \in (\mathscr{O} \times \mathscr{A})^{h-1}} \tilde{\pi}_h^k(a|\tau_{h-1}, o) \|\tilde{\pi}^k(\tau_{h-1}) \mathbb{Z}_h^{\theta^*\dagger} \mathbf{v}_{h-1}^{\theta^*}(\tau_{h-1})\|_1 \\
&= \sum_{(a,o) \in \mathscr{A} \times \mathscr{O}} \sum_{\tau_{h-1} \in (\mathscr{O} \times \mathscr{A})^{h-1}} \tilde{\pi}_h^k(a|\tau_{h-1}, o) \sum_{s_h \in \mathscr{S}} \mathbb{P}_{\theta^*}^{\tilde{\pi}^k}(\tau_{h-1}, s_h) \\
&= \sum_{o \in \mathscr{O}} \sum_{\tau_{h-1} \in (\mathscr{O} \times \mathscr{A})^{h-1}} \sum_{s_h \in \mathscr{S}} \mathbb{P}_{\theta^*}^{\tilde{\pi}^k}(\tau_{h-1}, s_h) = O.
\end{aligned}$$

We also have

$$\begin{aligned}
&\sum_{(a,o) \in \mathscr{A} \times \mathscr{O}} \sum_{\tilde{o} \in \mathscr{O}} \|w_{k,a,o,\tilde{o}}\|_1 \\
&= \sum_{s \in \mathscr{S}} \sum_{(a,o) \in \mathscr{A} \times \mathscr{O}} \left\| \frac{1}{2} [\mathbf{B}_h^{\theta^*}(a,o) - \mathbf{B}_h^{\hat{\theta}^k}(a,o)] \mathbb{Z}_h^{\theta^*} \mathbf{e}_s \right\|_1 \\
&\le \sum_{s \in \mathscr{S}} \frac{1}{2} \left(\sum_{(a,o) \in \mathscr{A} \times \mathscr{O}} \left\| \mathbf{B}_h^{\theta^*}(a,o) \mathbb{Z}_h^{\theta^*} \mathbf{e}_s \right\|_1 + \sum_{(a,o) \in \mathscr{A} \times \mathscr{O}} \left\| \mathbf{B}_h^{\hat{\theta}^k}(a,o) \mathbb{Z}_h^{\theta^*} \mathbf{e}_s \right\|_1 \right) \\
&\overset{(*)}{\le} \sum_{s \in \mathscr{S}} A \frac{\sqrt{S}}{\alpha} \|\mathbb{Z}_h^{\theta^*} \mathbf{e}_s\|_1 = \frac{S^{1.5} A}{\alpha}
\end{aligned}$$

where in (*) we have used Lemma E.4.

Applying Proposition B.2 with $\lambda = \frac{S}{\alpha^2}$, we have

$$\begin{aligned}
\text{(OP-REG-}h\text{)} &= \frac{1}{2}\sum_{k=1}^{K}\sum_{\tau_h}\tilde{\pi}^k(\tau_h)\|[\mathbf{B}_h^{\theta^*}(a_h,o_h) - \mathbf{B}_h^{\hat{\theta}^k}(a_h,o_h)]\mathbf{v}_{h-1}^{\theta^*}(\tau_{h-1})\|_1 \\
&= \sum_{k=1}^{K}\sum_{(a,o)\in\mathscr{A}\times\mathscr{O}}\sum_{\tilde{o}\in\mathscr{O}}\sum_{\tau_{h-1}\in(\mathscr{O}\times\mathscr{A})^{h-1}} |w_{k,a,o,\tilde{o}}^T x_{k,a,o,\tau_{h-1}}| \\
&\le \sqrt{\left(\frac{S}{\alpha^2} + \frac{4S}{\alpha^2}\beta\right) SAO^2K\log\left(1+SAO^2K\right)} \\
&\le 3\alpha^{-1}SO\sqrt{A\beta K\log\left(1+SAO^2K\right)}.
\end{aligned}$$

Using Step 1, we conclude that with probability at least $1-\frac{1}{K}$,

$$\begin{aligned}
\text{(TR-REG-II)} &\le \frac{\sqrt{S}}{\alpha}\left(\text{(OP-REG-0)} + \sum_{h=1}^{H-1}\text{(OP-REG-}h\text{)}\right) \\
&\le \frac{\sqrt{S}}{\alpha}\left(2\sqrt{\beta K} + (H-1)\cdot 3\alpha^{-1}SO\sqrt{A\beta K\log\left(1+SAO^2K\right)}\right) \\
&\le 3\alpha^{-2}HSO\sqrt{SA\beta K\log(1+SAO^2K)}.
\end{aligned}$$

Since TV distances are bounded by 1, we also have (TR-REG-II) $\le K$ with probability 1. Therefore, we have

$$\mathbb{E}\left[\text{(TR-REG-II)}\right] \le 3\alpha^{-2}HSO\sqrt{SA\beta K\log(1+SAO^2K)} + 1.$$

Finally, using Lemma D.2 and (18) we have

$$\begin{aligned}
\mathrm{BReg}(\phi^{\mathrm{PS4POMDP}}, K) &\le 2H + H\mathbb{E}\left[\text{(TR-REG)}\right] \\
&\le 3H + H\mathbb{E}\left[\text{(TR-REG-II)}\right] \\
&= \mathcal{O}\left(\alpha^{-2}H^2SO\sqrt{SA\beta K\log(1+SAO^2K)}\right) \\
&= \tilde{\mathcal{O}}\left(\alpha^{-2}H^2S^2O\sqrt{HA(SA+O)K}\right)
\end{aligned}$$

where in the last line we used the fact that $\beta = \tilde{\mathcal{O}}(\log(|\bar{\Theta}|)) = \tilde{\mathcal{O}}(HS^2A + HSO)$.

D.6 Proof of Corollary 4.6

Proof. All expectations and probabilities below are under $\hat{\phi}^{\mathrm{PS4POMDP}}$. Conditioned on $\mathcal{D}_k$, the distributions of θ^* and $\tilde{\theta}^k$ are ν^k and $\hat{\nu}^k$, respectively. Since $V_\theta^* \in [0,H]$, total variation and ϵ_k-optimality give

$$\begin{aligned}
\mathbb{E}[V_{\theta^*}^*] &\le \mathbb{E}[V_{\tilde{\theta}^k}^*] + H\mathbb{E}\left[\|\hat{\nu}^k - \nu^k\|_{\mathrm{TV}}\right] \\
&\le \mathbb{E}[V_{\tilde{\theta}^k}^{\tilde{\pi}^k}] + \epsilon_k + H\delta_k.
\end{aligned}$$

Using the notation of Appendix D, define

$$G_k := \left\{\theta\in\Theta : \ell(\theta,\mathcal{D}_k) \ge \bar{\ell}^*(\mathcal{D}_k) - \log(K|\bar{\Theta}|)\right\}.$$

Because G_k is determined by $\mathcal{D}_k$, another application of total variation and Lemma E.3 yields

$$\begin{aligned}
\mathbb{P}(\tilde{\theta}^k \notin G_k) &= \mathbb{E}[\hat{\nu}^k(G_k^c)] \\
&\le \mathbb{E}[\nu^k(G_k^c)] + \delta_k \\
&= \mathbb{P}(\theta^* \notin G_k) + \delta_k \le \frac{1}{K} + \delta_k.
\end{aligned}$$

Let $\vartheta^k = \iota(\tilde{\theta}^k)$. By Quick Fact 1, $\tilde{\theta}^k \in G_k$ implies $\vartheta^k \in \bar{\Theta}(\mathcal{D}_k)$. Proposition D.1 also gives

$$\left\| \mathbb{P}_{\tilde{\theta}^k}^{\tilde{\pi}^k} - \mathbb{P}_{\vartheta^k}^{\tilde{\pi}^k} \right\|_{\mathrm{TV}} \leq \frac{1}{K}.$$

Thus, by the triangle inequality and the fact that TV distances are at most one,

$$\begin{aligned}
\mathbb{E}\left[\left\| \mathbb{P}_{\tilde{\theta}^k}^{\tilde{\pi}^k} - \mathbb{P}_{\theta^*}^{\tilde{\pi}^k} \right\|_{\mathrm{TV}} \right] &\leq \mathbb{E}\left[\max_{\bar{\theta} \in \bar{\Theta}(\mathcal{D}_k)} \left\| \mathbb{P}_{\bar{\theta}}^{\tilde{\pi}^k} - \mathbb{P}_{\theta^*}^{\tilde{\pi}^k} \right\|_{\mathrm{TV}} \right] \\
&\quad + \mathbb{P}(\vartheta^k \notin \bar{\Theta}(\mathcal{D}_k)) + \frac{1}{K} \\
&\leq \mathbb{E}\left[\max_{\bar{\theta} \in \bar{\Theta}(\mathcal{D}_k)} \left\| \mathbb{P}_{\bar{\theta}}^{\tilde{\pi}^k} - \mathbb{P}_{\theta^*}^{\tilde{\pi}^k} \right\|_{\mathrm{TV}} \right] + \frac{2}{K} + \delta_k.
\end{aligned}$$

Combining these inequalities with

$$V_{\tilde{\theta}^k}^{\tilde{\pi}^k} - V_{\theta^*}^{\tilde{\pi}^k} \leq H \left\| \mathbb{P}_{\tilde{\theta}^k}^{\tilde{\pi}^k} - \mathbb{P}_{\theta^*}^{\tilde{\pi}^k} \right\|_{\mathrm{TV}}$$

and summing over episodes gives

$$\begin{aligned}
\mathrm{BReg}(\hat{\phi}^{\mathrm{PS4POMDP}}, K) &\leq 2H + H\mathbb{E}\left[\sum_{k=1}^{K} \max_{\bar{\theta} \in \bar{\Theta}(\mathcal{D}_k)} \left\| \mathbb{P}_{\bar{\theta}}^{\tilde{\pi}^k} - \mathbb{P}_{\theta^*}^{\tilde{\pi}^k} \right\|_{\mathrm{TV}} \right] \\
&\quad + \sum_{k=1}^{K} (\epsilon_k + 2H\delta_k).
\end{aligned}$$

The first two terms are the same as in Lemma 5.1. Their bounds in Appendices D.4 and D.5 use confidence-set guarantees valid under any learning algorithm and therefore apply unchanged. These give the general bound and, under Assumption 3.1, the weakly revealing bound, respectively. □

D.7 A guarantee for the exact posterior

Corollary D.4. *Let $C_K := 6\log(K|\bar{\Theta}|) + 8$, where $\bar{\Theta}$ is the finite surrogate set defined in Appendix D. Under any learning algorithm, for every $k \in [K]$, the exact posterior ν^k satisfies*

$$\mathbb{E}\left[\nu^k \left(\left\{ \theta \in \Theta : \sum_{j=1}^{k-1} \left\| \mathbb{P}_{\theta}^{\pi^j} - \mathbb{P}_{\theta^*}^{\pi^j} \right\|_{\mathrm{TV}}^2 \leq C_K \right\} \right) \right] \geq 1 - \frac{2}{K}.$$

Proof. Fix $k \in [K]$ and let A_k denote the set inside ν^k above. Define

$$G_k := \left\{ \theta \in \Theta : \ell(\theta, \mathcal{D}_k) \geq \bar{\ell}^*(\mathcal{D}_k) - \log(K|\bar{\Theta}|) \right\}.$$

Since G_k is determined by $\mathcal{D}_k$ and ν^k is the conditional law of θ^* given $\mathcal{D}_k$, Lemma E.3 gives

$$\mathbb{E}[\nu^k(G_k)] = \mathbb{P}(\theta^* \in G_k) \geq 1 - \frac{1}{K}.$$

Let $\mathcal{E}$ be the event in Lemma D.3, with $\mathbb{P}(\mathcal{E}) \geq 1 - 1/K$. For any $\theta \in G_k$, Quick Fact 1 implies $\iota(\theta) \in \bar{\Theta}(\mathcal{D}_k)$. On $\mathcal{E}$, Proposition D.1 and the triangle inequality therefore give

$$\begin{aligned}
\sum_{j=1}^{k-1} \left\| \mathbb{P}_{\theta}^{\pi^j} - \mathbb{P}_{\theta^*}^{\pi^j} \right\|_{\mathrm{TV}}^2 &\leq 2 \sum_{j=1}^{k-1} \left\| \mathbb{P}_{\iota(\theta)}^{\pi^j} - \mathbb{P}_{\theta^*}^{\pi^j} \right\|_{\mathrm{TV}}^2 + \frac{2(k-1)}{K^2} \\
&\leq 6\log(K|\bar{\Theta}|) + 6 + \frac{2(k-1)}{K^2} \leq C_K.
\end{aligned}$$

Thus $G_k \subseteq A_k$ on $\mathcal{E}$, and

$$\mathbb{E}[\nu^k(A_k)] \geq \mathbb{E}[\nu^k(G_k)] - \mathbb{P}(\mathcal{E}^c) \geq 1 - \frac{2}{K}.$$

□

This bound concerns trajectory distributions under the played policies and does not imply concentration on a unique model parameter, since hidden-state relabelings and unexplored components can prevent identification.

E Miscellaneous auxiliary results

Lemma E.1. *Let $\{Z_k\}_{k=1}^K$ be a real-valued random process adapted to the filtration $\{\mathcal{F}_k\}_{k=0}^K$. Suppose that $\psi_k(\lambda) = \log \mathbb{E}[\exp(\lambda Z_k)|\mathcal{F}_{k-1}]$ is well-defined for some $\lambda \in \mathbb{R}$. Then, for any $x > 0$,*

$$\mathbb{P}\left(\max_{k\in[K]} \sum_{j=1}^{k} [\lambda Z_j - \psi_j(\lambda)] < x\right) \geq 1 - e^{-x}$$

Proof. Define $M_0 = 1$ and $M_k = \exp\left(\sum_{j=1}^k [\lambda Z_j - \psi_j(\lambda)]\right)$, then

$$\begin{aligned}\mathbb{E}[M_k|\mathcal{F}_{k-1}] &= \exp\left(\sum_{j=1}^{k-1} [\lambda Z_j - \psi_j(\lambda)]\right) \cdot \mathbb{E}\left[\exp(\lambda Z_k - \psi_k(\lambda))|\mathcal{F}_{k-1}\right] \\ &= M_{k-1} \frac{\mathbb{E}[\exp(\lambda Z_k)|\mathcal{F}_{k-1}]}{\exp(\psi_k(\lambda))} = M_{k-1}.\end{aligned}$$

Therefore, M_k is a martingale with mean $\mathbb{E}[M_k] = \mathbb{E}[M_0] = 1$. By Doob's Martingale Inequality, we have

$$\mathbb{P}\left(\max_{k\in[K]} M_k \geq e^x\right) \leq \mathbb{E}[M_K]e^{-x} = e^{-x} \qquad \forall x > 0$$

The lemma is then established by taking the logarithm of both sides of the inequality in the condition. □

Lemma E.2. *Let $p, q \in \Delta_d$ be two probability distributions on the finite set $[d]$, then $\sum_{i=1}^d (\sqrt{p_i} - \sqrt{q_i})^2 \geq \|p - q\|_{\mathrm{TV}}^2$.*

Proof. We have

$$\begin{aligned}\|p - q\|_{\mathrm{TV}}^2 &= \frac{1}{4}\left(\sum_{i=1}^d |p_i - q_i|\right)^2 = \frac{1}{4}\left(\sum_{i=1}^d |(\sqrt{p_i} + \sqrt{q_i})(\sqrt{p_i} - \sqrt{q_i})|\right)^2 \\ &\leq \frac{1}{4}\left(\sum_{i=1}^d (\sqrt{p_i} + \sqrt{q_i})^2\right)\left(\sum_{i=1}^d (\sqrt{p_i} - \sqrt{q_i})^2\right) \\ &\leq \frac{1}{4}\left(\sum_{i=1}^d (2p_i + 2q_i)\right)\left(\sum_{i=1}^d (\sqrt{p_i} - \sqrt{q_i})^2\right) = \sum_{i=1}^d (\sqrt{p_i} - \sqrt{q_i})^2\end{aligned}$$

□

Lemma E.3. *Let $\bar{\Theta}$ be as specified in Proposition D.1 and $\ell, \bar{\ell}^*$ be as defined in* (12) *and* (13)*. For each $k \in [K]$, with probability at least $1 - \delta$, we have $\ell(\theta^*, \mathcal{D}_k) \geq \bar{\ell}^*(\mathcal{D}_k) - \log(|\bar{\Theta}|/\delta)$.*

Proof. For any $\bar{\theta} \in \bar{\Theta}$ and any $j \in [K]$, we have

$$\mathbb{E}\left[\frac{\mathbb{P}_{\bar{\theta}}^{\pi^j}(\tau^j)}{\mathbb{P}_{\theta^*}^{\pi^j}(\tau^j)} \,\middle|\, \mathcal{D}_j, \theta^*, \pi^j\right] = \sum_{\tilde{\tau}\in\mathscr{T}:\mathbb{P}_{\theta^*}^{\pi^j}(\tilde{\tau})>0} \mathbb{P}_{\theta^*}^{\pi^j}(\tilde{\tau}) \frac{\mathbb{P}_{\bar{\theta}}^{\pi^j}(\tilde{\tau})}{\mathbb{P}_{\theta^*}^{\pi^j}(\tilde{\tau})} \leq 1$$

Therefore,

$$\begin{aligned}&\mathbb{E}\left[\exp\left(\ell(\bar{\theta}, \mathcal{D}_k) - \ell(\theta^*, \mathcal{D}_k)\right)\right] \\ &= \mathbb{E}\left[\prod_{j=1}^{k-1} \frac{\mathbb{P}_{\bar{\theta}}^{\pi^j}(\tau^j)}{\mathbb{P}_{\theta^*}^{\pi^j}(\tau^j)}\right] = \mathbb{E}\left[\mathbb{E}\left[\prod_{j=1}^{k-1} \frac{\mathbb{P}_{\bar{\theta}}^{\pi^j}(\tau^j)}{\mathbb{P}_{\theta^*}^{\pi^j}(\tau^j)} \,\middle|\, \mathcal{D}_{k-1}, \theta^*\right]\right]\end{aligned}$$

$$= \mathbb{E}\left[\prod_{j=1}^{k-2}\frac{\mathbb{P}_{\bar{\theta}}^{\pi^j}(\tau^j)}{\mathbb{P}_{\theta^*}^{\pi^j}(\tau^j)}\mathbb{E}\left[\frac{\mathbb{P}_{\bar{\theta}}^{\pi^{k-1}}(\tau^{k-1})}{\mathbb{P}_{\theta^*}^{\pi^{k-1}}(\tau^{k-1})}\,\middle|\,\mathcal{D}_{k-1},\theta^*\right]\right] \le \mathbb{E}\left[\prod_{j=1}^{k-2}\frac{\mathbb{P}_{\bar{\theta}}^{\pi^j}(\tau^j)}{\mathbb{P}_{\theta^*}^{\pi^j}(\tau^j)}\right] \le \ldots \le 1$$

Using Markov Inequality, we have for any $\delta > 0$

$$\mathbb{P}\left(\ell(\bar{\theta},\mathcal{D}_k) - \ell(\theta^*,\mathcal{D}_k) \ge \log\left(\frac{1}{\delta}\right)\right) \le \delta.$$

Using Union Bound, we obtain

$$\mathbb{P}\left(\max_{\bar{\theta}\in\bar{\Theta}}\ell(\bar{\theta},\mathcal{D}_k) - \ell(\theta^*,\mathcal{D}_k) \ge \log\left(\frac{1}{\delta}\right)\right) \le |\bar{\Theta}|\delta.$$

The lemma is then obtained by replacing δ with $\delta/|\bar{\Theta}|$. □

Lemma E.4. *Assume that Assumption 3.1 is true and $\mathbf{B}_h^\theta$ is as defined in* (19)*. Let $x \in [0,1]^{\mathscr{A}\times\mathscr{O}}$ be such that $\sum_{a\in\mathscr{A}} x(a,o) = 1$ for all $o \in \mathscr{O}$. For any $\theta \in \Theta$ and any $\mathbf{w} \in \mathbb{R}^{\mathscr{O}}$, we have*

$$\sum_{a,o} x(a,o)\|\mathbf{B}_h^\theta(a,o)\mathbf{w}\|_1 \le \frac{\sqrt{S}}{\alpha}\|\mathbf{w}\|_1$$

Proof. Fix $s \in \mathscr{S}$, let $\mathbf{e}_s \in \mathbb{R}^{\mathscr{S}}$ be its corresponding indicator vector. Since $\mathbb{Z}_{h+1}^\theta \mathbb{T}_{h,a}^\theta \mathrm{diag}(\mathbb{Z}_h^\theta(o,\cdot))$ is a non-negative matrix, we have

$$\begin{aligned}
&\|\mathbb{Z}_{h+1}^\theta \mathbb{T}_{h,a}^\theta \mathrm{diag}(\mathbb{Z}_h^\theta(o,\cdot))\mathbf{e}_s\|_1 \\
&= \mathbf{1}^T \mathbb{Z}_{h+1}^\theta \mathbb{T}_{h,a}^\theta \mathrm{diag}(\mathbb{Z}_h^\theta(o,\cdot))\mathbf{e}_s \\
&= \mathbf{1}^T \mathbb{T}_{h,a}^\theta \mathrm{diag}(\mathbb{Z}_h^\theta(o,\cdot))\mathbf{e}_s = \mathbf{1}^T \mathrm{diag}(\mathbb{Z}_h^\theta(o,\cdot))\mathbf{e}_s = \mathbb{Z}_h^\theta(o,s)
\end{aligned}$$

Therefore, for any $\mathbf{w} \in \mathbb{R}^{\mathscr{O}}$

$$\begin{aligned}
&\sum_{a,o} x(a,o)\|\mathbf{B}_h^\theta(a,o)\mathbf{w}\|_1 = \sum_{a,o} x(a,o)\|\mathbb{Z}_{h+1}^\theta \mathbb{T}_{h,a}^\theta \mathrm{diag}(\mathbb{Z}_h^\theta(o,\cdot))(\mathbb{Z}_h^{\theta\dagger}\mathbf{w})\|_1 \\
&\le \sum_{a,o} x(a,o) \sum_{s\in\mathscr{S}} \left|[\mathbb{Z}_h^{\theta\dagger}\mathbf{w}](s)\right| \cdot \|\mathbb{Z}_{h+1}^\theta \mathbb{T}_{h,a}^\theta \mathrm{diag}(\mathbb{Z}_h^\theta(o,\cdot))\mathbf{e}_s\|_1 \\
&= \sum_{a,o} x(a,o) \sum_{s\in\mathscr{S}} \left|[\mathbb{Z}_h^{\theta\dagger}\mathbf{w}](s)\right| \cdot \mathbb{Z}_h^\theta(o,s) = \sum_o \left(\sum_a x(a,o)\right) \sum_{s\in\mathscr{S}} \left|[\mathbb{Z}_h^{\theta\dagger}\mathbf{w}](s)\right| \cdot \mathbb{Z}_h^\theta(o,s) \\
&= \sum_{s\in\mathscr{S}} \left|(\mathbb{Z}_h^{\theta\dagger}\mathbf{w})_s\right| \sum_o \mathbb{Z}_h^\theta(o,s) = \sum_{s\in\mathscr{S}} \left|[\mathbb{Z}_h^{\theta\dagger}\mathbf{w}](s)\right| = \|\mathbb{Z}_h^{\theta\dagger}\mathbf{w}\|_1 \le \frac{\sqrt{S}}{\alpha}\|\mathbf{w}\|_1
\end{aligned}$$

where we used Quick Fact 4 in the last inequality. □

Lemma E.5. *Assume that Assumption 3.1 is true and $\mathbf{B}_{H:h+1}^\theta$ is as defined in* (21)*. Fix $0 \le h < H$. For any policy $\pi \in \Pi$, any $\tau_h \in (\mathscr{O}\times\mathscr{A})^h$ such that $\pi(\tau_h) > 0$, any $\theta \in \Theta$, and any $\mathbf{w} \in \mathbb{R}^{\mathscr{O}}$, we have*

$$\sum_{\tau_{-h}\in(\mathscr{O}\times\mathscr{A})^{H-h}} \pi(\tau_{-h}|\tau_h)|\mathbf{B}_{H:h+1}^\theta(\tau_{-h})\mathbf{w}| \le \frac{\sqrt{S}}{\alpha}\|\mathbf{w}\|_1.$$

Proof. Fix $\tau_h \in (\mathscr{O}\times\mathscr{A})^h$. First, when $h = H-1$, we have $\mathbf{B}_{H:h+1}^\theta(\tau_{-h}) = \mathbf{e}_{o_H}^T$ and

$$\sum_{\tau_{-h}\in(\mathscr{O}\times\mathscr{A})^{H-h}} \pi(\tau_{-h}|\tau_h)|\mathbf{B}_{H:h+1}^\theta(\tau_{-h})\mathbf{w}| = \sum_{a_H,o_H} \pi_H(a_H|\tau_{H-1},o_H)|\mathbf{w}(o_H)| = \sum_{o_H}|\mathbf{w}(o_H)| = \|\mathbf{w}\|_1.$$

For the rest of the proof, assume $h < H-1$. First, note that

$$\mathbf{B}^{\theta}_{H:h+1}(\tau_{-h}) = \underbrace{\mathbf{e}^T_{o_H}\mathbb{Z}^{\theta}_H\mathbb{T}^{\theta}_{H-1,a_{H-1}}\mathrm{diag}(\mathbb{Z}^{\theta}_{H-1}(o_{H-1},\cdot))\cdots\mathbb{T}^{\theta}_{h+1,a_{h+1}}\mathrm{diag}(\mathbb{Z}^{\theta}_{h+1}(o_{h+1},\cdot))}_{=:\tilde{\mathbf{B}}^{\theta}_{H:h+1}(\tau_{-h})}\mathbb{Z}^{\theta\dagger}_{h+1}$$

For any $s_{h+1}\in\mathscr{S}$, we have

$$\pi(\tau_{-h}|\tau_h)|\tilde{\mathbf{B}}^{\theta}_{H:h+1}(\tau_{-h})\mathbf{e}_{s_{h+1}}| = \pi(\tau_{-h}|\tau_h)\mathbb{P}^{-}_{\theta}(\tau_{-h}|s_{h+1}) = \mathbb{P}^{\pi(\tau_h,\cdot)}_{\theta}(\tau_{-h}|s_{h+1})$$

where $\pi(\tau_h,\cdot)$ is the partial policy for timestamps $h+1$ to H obtained by fixing the historical trajectory τ_h.

Therefore, for any $\mathbf{w}\in\mathbb{R}^{\mathscr{O}}$,

$$\sum_{\tau_{-h}}\pi(\tau_{-h}|\tau_h)|\mathbf{B}^{\theta}_{H:h+1}(\tau_{-h})\mathbf{w}| = \sum_{\tau_{-h}}\pi(\tau_{-h}|\tau_h)|\tilde{\mathbf{B}}^{\theta}_{H:h+1}(\tau_{-h})(\mathbb{Z}^{\theta\dagger}_{h+1}\mathbf{w})|$$

$$\le\sum_{\tau_{-h}}\sum_{s_{h+1}\in\mathscr{S}}\pi(\tau_{-h}|\tau_h)|\tilde{\mathbf{B}}^{\theta}_{H:h+1}(\tau_{-h})\mathbf{e}_{s_{h+1}}|\cdot\left|(\mathbb{Z}^{\theta\dagger}_{h+1}\mathbf{w})_{s_{h+1}}\right| = \sum_{\tau_{-h}}\sum_{s_{h+1}\in\mathscr{S}}\mathbb{P}^{\pi(\tau_h,\cdot)}_{\theta}(\tau_{-h}|s_{h+1})\left|(\mathbb{Z}^{\theta\dagger}_{h+1}\mathbf{w})_{s_{h+1}}\right|$$

$$=\sum_{s_{h+1}\in\mathscr{S}}\left(\sum_{\tau_{-h}}\mathbb{P}^{\pi(\tau_h,\cdot)}_{\theta}(\tau_{-h}|s_{h+1})\right)\cdot\left|(\mathbb{Z}^{\theta\dagger}_{h+1}\mathbf{w})_{s_{h+1}}\right| = \|\mathbb{Z}^{\theta\dagger}_{h+1}\mathbf{w}\|_1 \le \frac{\sqrt{S}}{\alpha}\|\mathbf{w}\|_1$$

where we used Quick Fact 4 in the last inequality.

□

F Learning multi-agent POMDPs

In this appendix, we extend our results to learning problems on multi-agent POMDPs (MA-POMDPs), which are models that involve multiple learning agents with the same objective but different information about the system. A multiple-agent POMDP can be characterized by a tuple $(I,\mathscr{S},\mathscr{A},\mathscr{O},H,b_1,T,Z,r)$, where $I\in\mathbb{N}$ is the number of agents; $\mathscr{S}$ is the state space with $|\mathscr{S}|=S$; $\mathscr{A}=\prod_{i=1}^{I}\mathscr{A}^i$ is the joint action space with $|\mathscr{A}|=A$; $\mathscr{O}=\prod_{i=1}^{I}\mathscr{O}^i$ is the set of joint observations with $|\mathscr{O}|=O$; $H\in\mathbb{N}$ is the horizon length; $b_1\in\Delta(\mathscr{S})$ is the distribution of the initial state; $T=(T_h)_{h=1}^{H-1}, T_h:\mathscr{S}\times\mathscr{A}\mapsto\Delta(\mathscr{S})$ is the state transition kernel; $Z=(Z_h)_{h=1}^{H}, Z_h:\mathscr{S}\mapsto\Delta(\mathscr{O})$ is the joint observation kernel; $r=(r_h)_{h=1}^{H}, r_h:\mathscr{O}\times\mathscr{A}\mapsto[0,1]$ is the instantaneous reward function.

We assume that $I,\mathscr{S},\mathscr{A},\mathscr{O},H,r$ are known to the learning agents. The quantities b_1,T and Z are (in general) unknown to the agents. In the same way as in Section 3, we assume that b_1,T,Z are parameterized by a parameter $\theta\in\Theta$. The parameter θ has a prior distribution $\nu^1\in\Delta(\Theta)$.

In one episode, agent i's individual policy is characterized by a collection of mappings $\pi_{\cdot;i}=(\pi_{h;i})_{h=1}^{H}, \pi_{h;i}:(\mathscr{O}^i\times\mathscr{A}^i)^{h-1}\times\mathscr{O}^i\mapsto\mathscr{A}^i$, that depends only on its individual information. A joint policy $\pi=(\pi_{\cdot;i})_{i\in[I]}$ consists of the individual policies of all agents. Let $\mathit{\Pi}$ denote the space of all deterministic joint policies.

At the beginning of each learning episode, the learning agents share all of their action and observation history with each other. Each agent then picks an individual policy based on the collective history (with possible randomization over policies) and uses it during the episode. We assume that the agents have access to a common randomness source, and hence their random policy choices can be correlated. The Bayesian regret is defined in the same way as in Section 3, where we compare the expected total reward under the given (joint) learning algorithm against the best joint policy with respect to the true MA-POMDP parameter.

The Posterior Sampling for Multi-Agent POMDPs (PS4MAPOMDP) algorithm works in a similar way to its single-agent counterpart: At the beginning of each episode k, a common sample $\tilde{\theta}^k$ is drawn based on the latest posterior distribution based on the *collective* action and observation history, then the agents collectively invoke an MA-POMDP solving oracle to obtain a joint policy $\tilde{\pi}^k$. Note

Algorithm 2 PS4MAPOMDP: Agent i

Input: Prior $\nu^1 \in \Delta(\Theta)$; Number of episodes K
for $k = 1$ to K **do**
  Use common randomness source to sample $\tilde{\theta}^k \sim \nu^k$
  Invoke the MA-POMDP solving oracle to obtain a policy $\tilde{\pi}^k \in \arg\max_{\pi \in \Pi}(V^{\pi}_{\tilde{\theta}^k})$
  Apply $\tilde{\pi}^k_{:,i}$ in the k-th episode
  At the end of k-th episode, Share the local trajectory $\tau^{k,i} = (a^k_{1:H;i}, o^k_{1:H;i})$ with other agents
  Use $\tau^k = (\tau^{k,j})_{j \in [I]}$ to compute new posterior $\nu^{k+1} \in \Delta(\Theta)$ using (3)
end for

that the above steps can be done separately by each agent using the common randomness source. Then each agent $i \in [I]$ uses its individual policy $\tilde{\pi}^k_{:,i}$ to take an action during the episode.

Under the assumption of action and observation sharing at the beginning of each episode, the MA-POMDP learning problem can be seen as a single-agent problem where the learning agent knows all the actions and observations but is artificially restricted to use only policies from Π, the set of joint policies. Per Remark 4.5, we obtain the following result.

Corollary F.1. *The Bayesian regret of the PS4MAPOMDP algorithm ϕ applied to a general MA-POMDP learning problem satisfies*

$$\mathrm{BReg}(\phi, K) \leq \tilde{\mathcal{O}}\left(H^2\sqrt{(S^2A + SO)(OA)^H K}\right).$$

Furthermore, if Assumption 3.1 holds for the joint observation kernel Z, then

$$\mathrm{BReg}(\phi, K) \leq \tilde{\mathcal{O}}\left(\alpha^{-2}H^2S^2O\sqrt{HA(SA + O)K}\right).$$

Note that Assumption 3.1 in multi-agent POMDP setting does not imply that individual observations are necessarily informative to a degree. It only requires the *joint observation* to be informative. A special case that satisfies Assumption 3.1 is the Dec-MDP model where the joint observation uncovers the underlying state perfectly.

G Experimental details

We use the NUTS implementation in NumPyro [Phan et al., 2019] under Apache-2.0 license. ULMC is adapted from [Ishfaq et al., 2024] under MIT license. We use the official SARSOP implementation, primarily licensed under GPL v2.

G.1 PPO baselines

In total, we consider eight PPO variants adapted from RLBenchNet [Smirnov and Gu, 2025], covering feedforward, recurrent, attention-based, and state-space memory architectures, as summarized in Table 2.[7] Since this work focuses on regret-minimizing algorithms for POMDPs, the PPO variants are included only as baselines, and we refer readers to prior work for detailed discussions of PPO training and memory architectures.

For reproducibility, we summarize the common training hyperparameters in Table 3, and report model-specific hyperparameters for the PPO variants in Tables 4 and 5. All hyperparameters are adapted from the experiments on Classic Control tasks in [Smirnov and Gu, 2025].

[7] PPO and its memory-augmented variants are adapted from [Smirnov and Gu, 2025] under MIT license.

Table 2: PPO baselines and their policy architectures.

Method	Policy Architecture	Reference
PPO	Feedforward MLP	Schulman et al. [2017]
PPO-Stack4	MLP with Atari-style stacked observations (4 steps)	Schulman et al. [2017]
PPO-LSTM	Long Short-Term Memory (LSTM)	Hochreiter and Schmidhuber [1997]
PPO-GRU	Gated Recurrent Unit (GRU)	Cho et al. [2014]
PPO-TrXL	Transformer-XL (TrXL)	Dai et al. [2019]
PPO-GTrXL	Gated Transformer-XL (GTrXL)	Parisotto et al. [2020]
PPO-Mamba	Mamba state-space model	Gu and Dao [2024]
PPO-Mamba2	Mamba-2 state-space model	Dao and Gu [2024]

Table 3: Common PPO Training Hyperparameters by Environment

Parameter	Tiger	RiverSwim	Random-S10A4	Sparse-S50A4
# Environments	4	4	4	4
# Steps / env (same as H)	10	40	20	50
Batch size	40	160	80	200
Mini-batch size (same as H)	10	40	20	50
Update epochs	4	4	4	4
Discount factor (γ)	0.99	0.99	0.99	0.99
GAE λ	0.95	0.95	0.95	0.95
Learning rate (Adam)	2.5×10^{-4}	2.5×10^{-4}	2.5×10^{-4}	2.5×10^{-4}
Entropy coefficient	0.01	0.01	0.01	0.01
Value loss coef. (c_v)	0.5	0.5	0.5	0.5
Max grad norm	0.5	0.5	0.5	0.5

G.2 Tiger

We consider the following POMDP adapted from the widely used Tiger example [Cassandra et al., 1994]. In this POMDP, a contestant is faced with two doors. Behind one door is a tiger and behind the other door is a treasure. The contestant at each time faces three choices: opens either door and end the game, or listen (for the roar of the tiger) and defer the decision. However, listening is not accurate and has a cost. Unlike the classical infinite horizon version of Tiger, we consider a version with a finite horizon H, and the game ends instead of resets after the contestant opens a door. The experimental results are based on the version where $H = 10, \beta = 0.99$. The full description $(\mathscr{S}, \mathscr{A}, \mathscr{O}, H, b_1^\theta, T^\theta, Z^\theta, r)$ is provided as follows:

- **State space:** $\mathscr{S} = \{\text{TL}, \text{TR}, \text{CD}, \text{CA}, \text{End}\}$ (tiger left; tiger right; contestant dead; contestant alive; game over).
- **Action space:** $\mathscr{A} = \{\text{listen}, \text{OL}, \text{OR}\}$ (open left; open right)
- **Observation space:** $\mathscr{O} = \{\text{HL}, \text{HR}, \text{CD}, \text{CA}, \text{End}\}$ (hear left; hear right; contestant dead; contestant alive; game over)
- **Parameter to be learned:** A single parameter $\theta \in [0, 0.5]$ for the observation kernel. The initial distribution and transition kernels are known under this parameterization.
- **Initial state distribution:** $b_1(\text{TL}) = b_1(\text{TR}) = 0.5$
- **Transition:** $\mathbb{T}_h(s'|s, a) = \mathbb{T}(s'|s, a)$ is stationary and deterministic
  - End is an absorbing state, i.e. $\mathbb{T}(\text{End}|\text{End}, *) = 1$
  - CD or CA always leads to End: $\mathbb{T}(\text{End}|\text{CD}, *) = \mathbb{T}(\text{End}|\text{CA}, *) = 1$
  - $\mathbb{T}(\text{TL}|\text{TL}, \text{listen}) = \mathbb{T}(\text{TR}|\text{TR}, \text{listen}) = 1$
  - $\mathbb{T}(\text{CD}|\text{TL}, \text{OL}) = \mathbb{T}(\text{CD}|\text{TR}, \text{OR}) = 1$
    $\mathbb{T}(\text{CA}|\text{TR}, \text{OL}) = \mathbb{T}(\text{CA}|\text{TL}, \text{OR}) = 1$
- **Observation Kernel:** $\mathbb{Z}_h(o|s) = \mathbb{Z}(o|s)$ is stationary
  - $\mathbb{Z}(\text{HL}|\text{TL}) = \mathbb{Z}(\text{HR}|\text{TR}) = 0.5 + \theta$ for $h \in [H]$

Table 4: Model-Specific Hyperparameters for Feedforward and RNN-Based PPO Variants

Parameter	**PPO**	**PPO-Stack4**	**PPO-LSTM**	**PPO-GRU**
Hidden dimension	256	256	256	256
Observation stack	1	4	–	–
RNN type	–	–	LSTM	GRU
RNN hidden dimension	–	–	128	160

Table 5: Model-Specific Hyperparameters for Attention and State-Space PPO Variants

Parameter	**PPO-TrXL**	**PPO-GTrXL**	**PPO-Mamba**	**PPO-Mamba2**
Hidden dimension	284	336	284	160
Transformer dimension	284	336	–	–
# Transformer layers	1	2	–	–
# Attention heads	2	2	–	–
Memory length	32	32	–	–
Mamba version	–	–	v1	v2
State dimension (d_{state})	–	–	64	128
Convolution width (d_{conv})	–	–	4	4
Expansion ratio	–	–	2	2
Mamba learning rate	–	–	1.5×10^{-4}	1.5×10^{-4}

  - $\mathbb{Z}(\text{HR}|\text{TL}) = \mathbb{Z}(\text{HL}|\text{TR}) = 0.5 - \theta$ for $h \in [H]$
  - $\mathbb{Z}(\text{CD}|\text{CD}) = \mathbb{Z}(\text{CA}|\text{CA}) = \mathbb{Z}(\text{End}|\text{End}) = 1$
- **Reward function:** $r_h(o, a), h \in [H]$

$$r_h(o,a) = -100\beta^{h-1}\mathbf{1}_{\{o=\text{CD}\}} + 10\beta^{h-1} \cdot \mathbf{1}_{\{o=\text{CA}\}} - 1 \cdot \beta^h \cdot \mathbf{1}_{\{a=\text{listen}\}}$$

  where $\beta \in (0, 1)$ is a discount factor.

Note that the action of opening door at time H bears no consequence in this model.

Results. Figure 3 shows that PS4POMDP exhibits sublinear cumulative regret across all three Tiger instances. When normalized by K, the regret converges to zero, and when normalized by $\sqrt{K}$, it approaches a constant. In contrast, PPO-based baselines incur substantially larger regret, and their regret normalized by $\sqrt{K}$ continues to increase with K.

Figure 4 provides a more comprehensive look at how PS4POMDP learns the underlying POMDP with $\theta^* = 0.3$. We see that after some initial warm-up, the algorithm is able to quickly learn the underlying parameter θ and the posterior samples $\tilde{\theta}^k$ (shown on the right plot in Figure 4) become close to the true θ. As a result, the policy obtained from the POMDP solver is nearly (or exactly) optimal.

OMLE comparison on discretized Tiger

As discussed in Section 6, the natural empirical competitors of PS4POMDP are other provably efficient online POMDP-learning algorithms. We focus here on Optimistic Maximum Likelihood Estimation (OMLE) Liu et al. [2022], the closest model-based competitor in the literature. OMLE maintains a confidence set of plausible models that fit the observed trajectories well in likelihood, and at each episode plays the policy that achieves the highest value over any model in this set. Each episode therefore requires a joint optimization over models and policies, which is intractable in general.

Discretized setting. To enable a direct comparison, we restrict the problem to a one-dimensional family of Tiger POMDPs. The unknown parameter is the observation-noise parameter θ from Section 6; we discretize the interval into $M = 200$ candidate values uniformly spaced on $(0, 0.5]$,

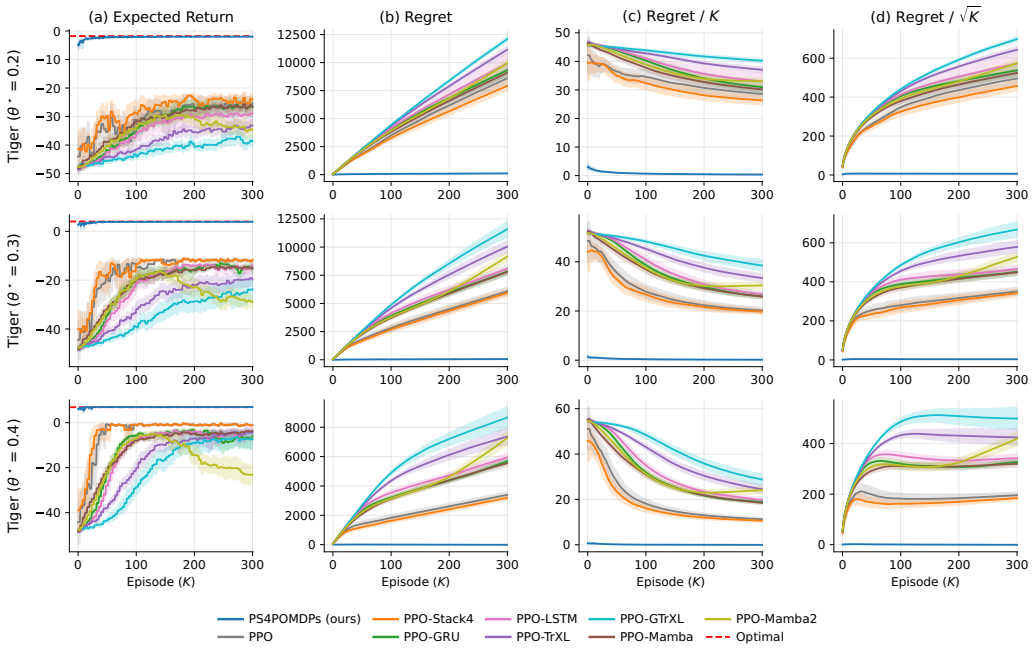


Figure 3: Experimental results on three *Tiger* instances with $\theta^\star \in \{0.2, 0.3, 0.4\}$. **(a)** Expected return, **(b)** cumulative regret, **(c)** regret divided by K, and **(d)** regret divided by $\sqrt{K}$. Shaded regions denote 95% bootstrapped confidence intervals over 20 independent runs.

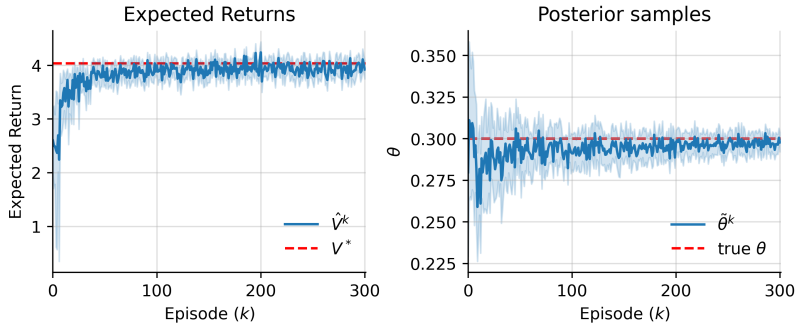


Figure 4: Detailed results for PS4POMDP in *Tiger* with $\theta^* = 0.3$. **Left**: Expected returns for PS4POMDP policies (estimated with 1,000 rollout trajectories). **Right**: Posterior samples of PS4POMDP with the single-parameter model. Shaded regions denote 95% bootstrapped confidence intervals over 20 independent runs.

denoted Θ_M. Each candidate $\theta \in \Theta_M$ defines a fully specified Tiger POMDP, and for each such POMDP we precompute a near-optimal policy by running SARSOP on the corresponding model. With both candidate models and policies enumerated in advance, we approximate OMLE's joint maximization by a finite search over precomputed model-policy pairs:

$$(\hat{\theta}^k, \hat{\pi}^k) \;=\; \arg \max_{(\theta,\pi)\in\{(\theta,\pi^\star_\theta):\theta\in\Theta_M\}} V^\pi_\theta \text{ subject to } \quad -\log p_\theta(\tau_{1:k-1}) \;\le\; -\log p_{\hat{\theta}^{k-1}_{\text{MLE}}}(\tau_{1:k-1}) + \beta,$$

where $\Pi_M = \{\pi^\star_\theta : \theta \in \Theta_M\}$ is the precomputed policy set and $\beta \ge 0$ is the confidence-radius hyperparameter from Liu et al. [2022].

Algorithms compared. We compare three OMLE configurations with different confidence radii ($\beta \in \{0.1, 1, 5\}$) against PS4POMDP with a uniform prior over Θ_M. All algorithms share the

same enumerated policy set Π_M, so the comparison isolates the effect of the model-update strategy (optimistic confidence set vs. posterior sampling).

Results. Figure 5 reports expected return and regret metrics on Tiger with $\theta^\star = 0.2$ and $\theta^\star = 0.3$ over 300 episodes, averaged across multiple repeats. PS4POMDP is competitive with the best OMLE configuration while avoiding the need to tune a confidence radius β. The OMLE curves are sensitive to this choice: small β shrinks the confidence set quickly but risks excluding the true model, while large β admits too many candidates and slows learning. We emphasize that this comparison is only feasible because the discretized parameter space is one-dimensional; in the full continuous-parameter or general-POMDP setting, OMLE's joint optimization over models and policies is no longer tractable, which is why this is the only direct comparison reported in this paper.

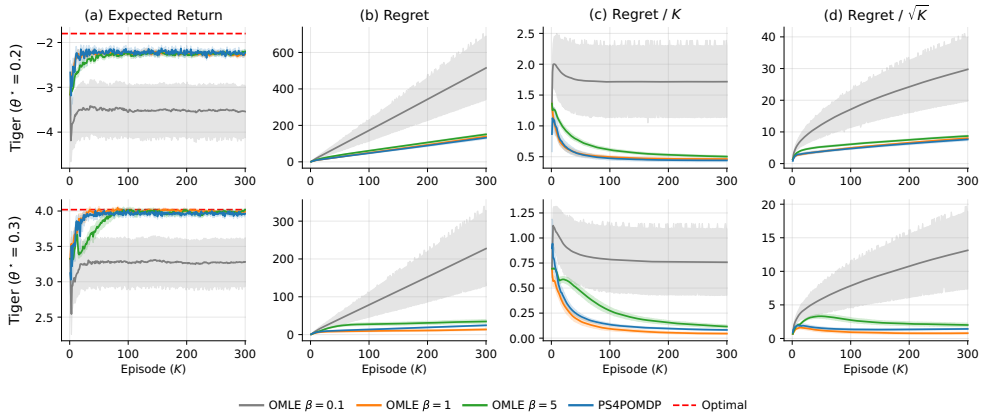


Figure 5: Discretized *Tiger* comparison with $\theta^\star \in \{0.2, 0.3\}$. **(a)** Expected return, **(b)** cumulative regret, **(c)** regret divided by K, and **(d)** regret divided by $\sqrt{K}$. Lines show means with shaded 95% confidence intervals across 50 repeats.

G.3 RiverSwim

RiverSwim (POMDP) formulation. We formulate the RiverSwim problem [Strehl and Littman, 2008, Osband et al., 2013] as a POMDP by introducing sensor errors. A RiverSwim POMDP instance with L states is defined by the following components:

- **State space:** $\mathscr{S} = \{1, 2, \ldots, L\}$, where L is the length of the river.
- **Action space:** $\mathscr{A} = \{\text{Left}, \text{Right}\}$.
- **Observation space:** $\mathscr{O} = \{1, 2, \ldots, L\}$.
- **Initial state distribution:** $b_1(1) = 1$, i.e., $s_1 = 1$ almost surely.
- **Transition kernel:** $T_h(s'|s, a) = T(s'|s, a)$ is stationary, as described in the transition diagram (Figure. 6).

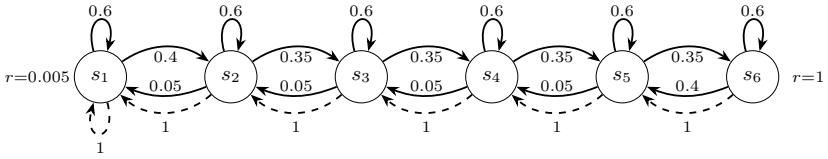


Figure 6: Transition diagram for RiverSwim with $L = 6$. Continuous and dashed arrows correspond to actions "Right" and "Left", respectively. (Modified from Osband et al. [2013])

- **Observation kernel:** $Z_h(o|s) = Z(o|s)$ is stationary with parameters $q = (q_1, q_2, q_3) \in \Delta(3)$:
  - $Z(1|1) = 1 = Z(L|L)$
  - $Z(2|2) = Z(L-1|L-1) = q_1$
  - $Z(3|2) = 1 - q_1 = Z(L-2|L-1)$
  - For $s = 3, ..., L-2$:

$$Z(o|s) = \begin{cases} q_1 & \text{if } o = s, \\ q_2 & \text{if } o = s+1, \\ q_3 & \text{if } o = s-1. \end{cases}$$

  In other words, the algorithm may encounter a sensor error and miscalculate it's distance to the target state L when it is in the middle of the river. However, it sees its true position when it is at the starting state 1 or when it reaches the target state L.

- **Reward function:** $r_h(o, a) = r(o, a)$ is stationary and depends only on the observation:

$$r(o, a) = \frac{5}{1000} \cdot \mathbf{1}_{\{o=1\}} + 1 \cdot \mathbf{1}_{\{o=L\}}$$

Simulated problem instance. For experiment, we use a RiverSwim POMDP instance with $L = 6$, $p = (0.6, 0.35, 0.05)$, and $q = (0.6, 0.2, 0.2)$ and horizon $H = 40$. Without the observation kernel, this environment setup is identical to RiverSwim MDP considered by Osband et al. [2013]. Whereas, in our POMDP environment, an algorithm will acquire an incorrect estimate of its location (i.e., state) 40% of the time when it is in the middle of the river.

Results. Figure 7 reports the RiverSwim learning curves. PS4POMDP identifies high-value policies after relatively few episodes and keeps cumulative regret substantially below the PPO baselines. The PPO variants eventually improve over random behavior, but their learning curves remain well below the SARSOP reference value and their normalized regret continues to grow over the plotted horizon. This is consistent with RiverSwim's role as a hard-exploration benchmark: the challenge is reaching and exploiting the far-right reward state, rather than merely retaining a longer observation history.

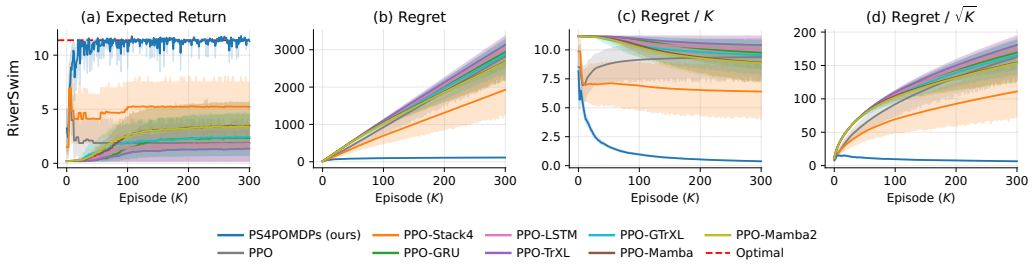


Figure 7: Experimental results on *RiverSwim*. **(a)** Expected return, **(b)** cumulative regret, **(c)** regret divided by K, and **(d)** regret divided by $\sqrt{K}$. Shaded regions denote 95% bootstrapped confidence intervals over 20 independent runs.

G.4 Random-S10A4

We conduct additional experiments on randomly generated POMDPs (*Random-S10A4*) derived from RiverSwim. Each *Random-S10A4* instance has horizon $H = 20$ with $S = 10$, $A = 4$, and $O = 10$, and shares the same observation kernel and reward model as RiverSwim. The transition kernel is stationary and randomly sampled from the uniform prior over transition kernels. Figure 8 shows that the same pattern persists across five random instances: PS4POMDP reaches higher expected returns and incurs lower regret than the PPO baselines, while the PPO variants often settle at visibly suboptimal policies.

G.5 Sparse-S50A4

The *Sparse-S50A4* family is a medium-scale extension of the *Random-S10A4* environments introduced in Section 6 and Appendix G.4. We use it to demonstrate that the PS4POMDP pipeline continues to work on substantially larger tabular POMDPs once the inference subroutine is replaced with a gradient-based posterior sampler.

Environment. A *Sparse-S50A4* instance has horizon $H = 50$, state space $S = 50$, action space $A = 4$, and observation space $O = 50$. The transition kernel is sampled from a sparse Dirichlet prior: for each (s, a), the support of $T(\cdot \mid s, a)$ is restricted to a small randomly-chosen subset of next states, and probabilities on the support are drawn from a Dirichlet distribution. The boundary states $s = 0$ and $s = S - 1$ are fully observable. For interior states the observation kernel returns the true state with probability $1 - p_{\text{noise}}$ and otherwise spreads the remaining mass uniformly over the other observations,

$$Z(o \mid s) = (1 - p_{\text{noise}})\,\mathbf{1}[o = s] + \frac{p_{\text{noise}}}{O - 1}\,\mathbf{1}[o \neq s], \qquad s \in \{1, \ldots, S - 2\},$$

with $p_{\text{noise}} = 0.5$ in our runs. The reward function is inherited from the *RiverSwim*-style construction in Appendix G.3, scaled to the larger state and observation spaces. We report results on the inst-0 instance in the main figure; the remaining instances yield qualitatively similar curves.

Posterior sampling via ULMC. We use underdamped Langevin Monte Carlo (ULMC), adapted from Ishfaq et al. [2024]. Sampler hyperparameters are listed in Table 6.

Update schedule. The posterior sample and the SARSOP policy are refreshed every 4 episodes, matching the update frequency of the PPO baselines. The first 50 episodes use random actions to warm-start the sampler.

Per-instance results. Figure 9 reports the four learning-curve metrics across all five *Sparse-S50A4* instances. The qualitative pattern matches the main figure (instance 0): PS4POMDP attains the highest expected return and the lowest normalized regret on every instance, while the PPO variants plateau visibly below $V^\star$.

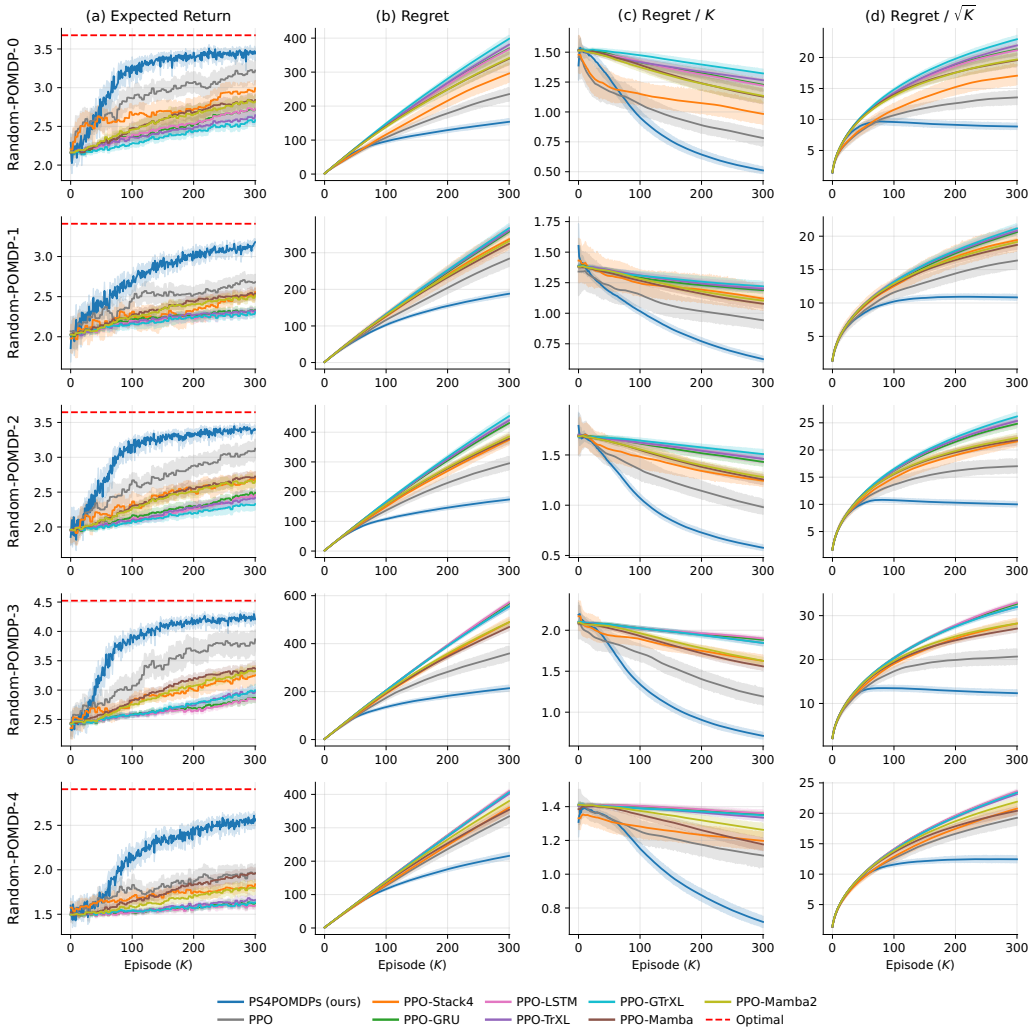


Figure 8: Experimental results on five *Random-S10A4* instances. **(a)** Expected return, **(b)** cumulative regret, **(c)** regret divided by K, and **(d)** regret divided by $\sqrt{K}$. Shaded regions denote 95% bootstrapped confidence intervals over 20 independent runs.

Table 6: *Sparse-S50A4* hyperparameters and ULMC sampler configuration.

Component	**Setting**
Dimensions	$S = 50$, $A = 4$, $O = 50$, $H = 50$
Initial state	$s_0 = 0$ (deterministic)
Transition kernel	Dirichlet($\mathbf{1}_{25}$) over 25 randomly sampled next-states per (s, a)
Observation kernel	Boundary states ($s \in \{0, S-1\}$) deterministic; interior states emit true state w.p. $1 - p_{\text{noise}}$, else uniform over remaining $O - 1$ obs ($p_{\text{noise}} = 0.5$)
Reward	$r_{\text{start}} = 0.005$ at $o = 0$, $r_{\text{goal}} = 1$ at $o = O-1$ (*RiverSwim*-style)
Posterior sampler	EulerULMC (PyTorch)
ULMC steps per update	100
Base step size η_0	0.03
Step-size decay	$\eta_k = \eta_0 / k^{0.25}$ across episodes k
Noise scale	0.1
Batch	full trajectory buffer (no minibatching)
Warm-up episodes (random actions)	50
Model and policy update frequency	every 4 episodes
Total episodes per run	1000
Planner	SARSOP
Discount γ	0.99
Per-call timeout	60 s
Target precision	0.05
Evaluation	1,000 independent rollouts every 10 episodes
Seeds per instance	10

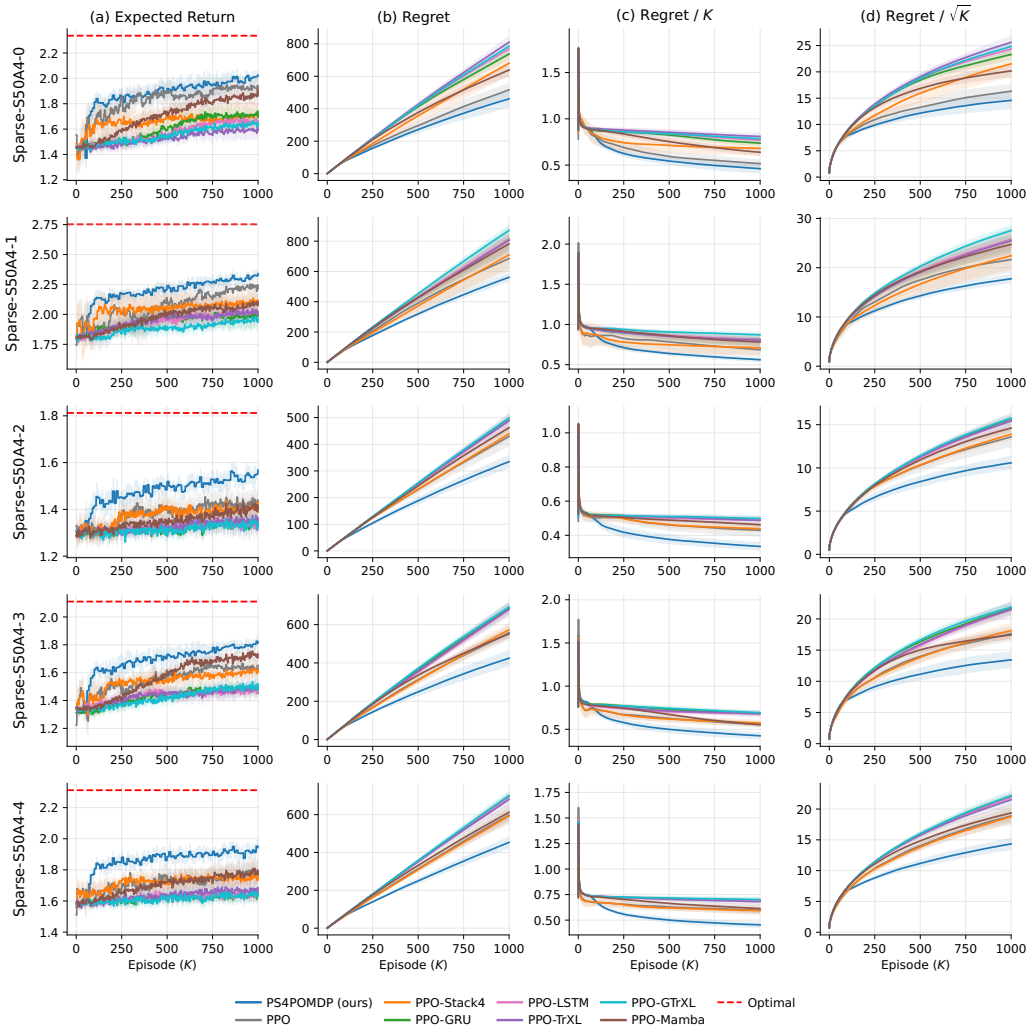


Figure 9: Experimental results on five *Sparse-S50A4* instances. **(a)** Expected return, **(b)** cumulative regret, **(c)** regret divided by K, and **(d)** regret divided by $\sqrt{K}$. The expected return of a policy is evaluated via $1{,}000$ Monte Carlo rollouts. Shaded regions denote 95% confidence intervals over 10 independent seeds.

G.6 Online-planner ablation

Learning protocol. Table 1 compares four planning subroutines over $K = 200$ learning episodes, with 10 seeds per configuration. RiverSwim and Random-S10A4 use one NUTS chain with 20 warmup steps and 20 sampling steps per update. Sparse-S50A4 uses the EulerULMC configuration in Appendix G.5, including 50 initial random-action episodes and model and policy updates every four episodes. The SARSOP result on Sparse-S50A4 is computed from the first 200 episodes of the existing instance-0 runs.

We use the public implementations of POMCPOW and AdaOPS in Julia and DESPOT in C++, integrated with the Python learning loop. POMCPOW uses 3,000 tree queries per action. DESPOT uses 500 particles and a 0.5-second planning budget per action; AdaOPS also uses a 0.5-second budget per action. SARSOP, POMCPOW, and DESPOT use planning discount $\gamma = 0.99$, while AdaOPS uses $\gamma = 1$. Reported returns are undiscounted environment returns. For the Sparse-S50A4 POMCPOW and AdaOPS adapters, the action-independent observation reward is converted to its state-conditional mean $\bar{r}(s, a) = \sum_o Z(o \mid s) r(o, a)$.

Evaluation protocol. For every column of Table 1, cumulative regret is estimated as $K\widehat{V}^\star - \sum_{k=1}^{K} R_k$, where R_k is the realized return in learning episode k and $\widehat{V}^\star$ is the SARSOP-on-true-model reference. Separate policy-evaluation rollouts would require additional action searches for the online planners, so this table uses realized returns for both PS4POMDP and PPO. The Best PPO variants are Mamba on RiverSwim, Mamba2 on Random-S10A4, and GRU on Sparse-S50A4, using their existing learning-return logs. Each cell reports the mean and 95% CI over 10 seeds; the comparisons are between means and do not imply statistically significant differences in every cell.

Planning on the true models. Table 7 evaluates the planners when the true model is given, separating planning performance from learning. Each cell uses 200 rollouts. SARSOP, POMCPOW, and AdaOPS use paired environment seeds; DESPOT uses its C++ simulator with unpaired rollouts. On RiverSwim, AdaOPS and SARSOP obtain identical realized returns on the 200 paired rollouts. The planner budgets are those listed above.

Table 7: True-model planner comparison: undiscounted return (higher is better), mean $\pm$ 95% CI over 200 rollouts.

Environment	SARSOP	POMCPOW	DESPOT	AdaOPS
RiverSwim	11.570 ± 0.637	9.985 ± 0.634	11.271 ± 0.646	11.570 ± 0.637
Random-S10A4	3.812 ± 0.272	3.619 ± 0.275	3.518 ± 0.237	3.365 ± 0.267
Sparse-S50A4	2.365 ± 0.212	2.189 ± 0.201	2.144 ± 0.192	1.974 ± 0.204

G.7 Sensitivity to the NUTS budget

We vary the number of NUTS sampling steps $m \in \{5, 20, 100, 1000\}$ per posterior update. Each update uses one chain with m warmup and m sampling steps; the final sample is used for planning. Figure 10, left, shows similar regret across the tested 200-fold range, with overlapping confidence intervals. Both environments use $K = 100$ learning episodes and 10 seeds per configuration. Table 8 reports cumulative regret based on separate Monte Carlo policy evaluations, as in the main learning curves. This measures sensitivity to the inference budget; it does not directly measure the total-variation distance between the approximate and exact parameter posteriors.

Table 8: NUTS-budget sensitivity at $K = 100$: mean cumulative regret $\pm$ 95% CI over 10 seeds.

Environment	$m = 5$	$m = 20$	$m = 100$	$m = 1000$
RiverSwim	115.7 ± 7.1	108.0 ± 8.2	109.8 ± 5.6	115.4 ± 9.7
Random-S10A4	120.4 ± 6.1	115.5 ± 6.8	113.2 ± 9.6	116.9 ± 12.5

G.8 Prior-sensitivity protocol

Figure 10, right, holds the true model fixed and varies the symmetric Dirichlet concentration $\alpha_0 \in \{0.3, 0.5, 1, 3, 5\}$ for every row of both the transition and observation kernels. All other settings are fixed within each environment. We report cumulative regret at $K = 100$, using separate Monte Carlo policy evaluations and 10 seeds per concentration. The environments are RiverSwim and instance 0 of Random-S10A4. The error bars are the reported 95% confidence intervals, not uncertainty over a draw of the true model from the prior. Both environments have sparse observation kernels, even though Random-S10A4 has dense transitions; varying the joint prior therefore does not isolate transition-prior sensitivity.

To examine dependence on the prior on fixed problem instances, we vary the symmetric Dirichlet concentration α_0 for both transition and observation kernels. Figure 10, right, shows similar regret for $0.3 \leq \alpha_0 \leq 1$ and worse performance at larger concentrations. Larger α_0 concentrates the prior near uniform kernels, whereas both environments have sparse observation kernels and RiverSwim also has sparse transitions. The overlapping intervals for $0.3 \leq \alpha_0 \leq 1$ do not identify a uniquely best concentration. These fixed-instance results complement the Bayesian analysis; they do not establish a frequentist regret guarantee.

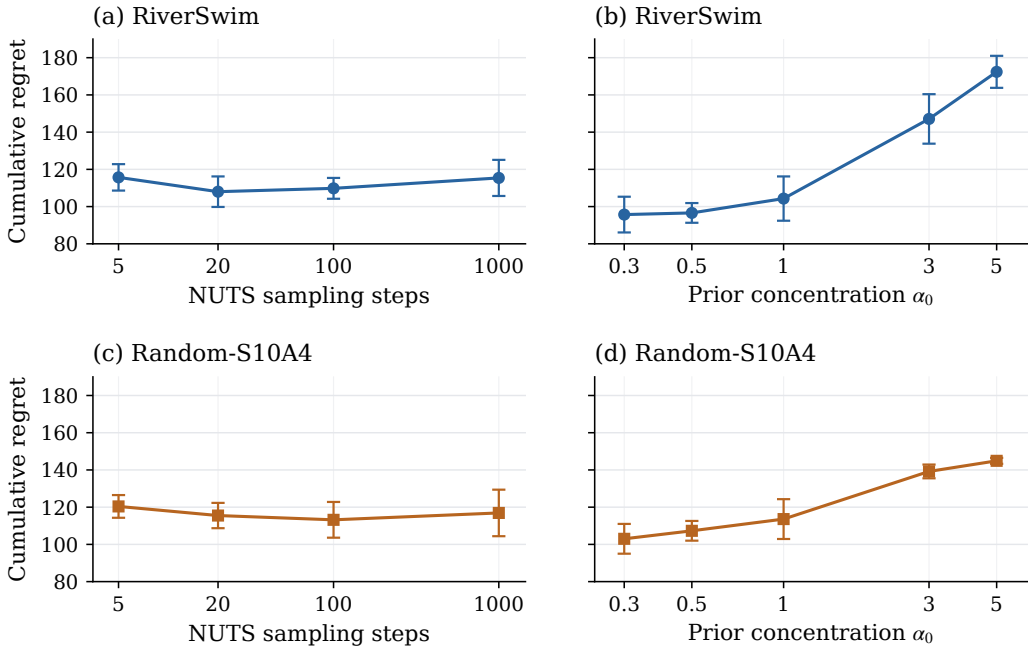


Figure 10: Sensitivity to posterior inference and the prior at $K = 100$: mean cumulative regret with 95% CIs over 10 seeds. **Top**: RiverSwim. **Bottom**: Random-S10A4. **Left**: varying NUTS sampling steps m. **Right**: varying the prior concentration α_0 for both transition and observation kernels. The two columns show separate experiments; see Appendices G.7 and G.8 for details.

G.9 Inference and planning times

Table 9 separates inference and planning time as the dataset grows on RiverSwim and Random-S10A4, where both transition and observation kernels are inferred. NUTS uses 1,000 warmup steps and 2,000 sampling steps; its timing blocks until the JAX computation finishes. Each NUTS cell averages five seeds. Each SARSOP cell averages 50 posterior models (10 models from each of five seeds), with a 120-second solver timeout, target precision 0.05, and planning discount $\gamma = 0.999$. All measurements were collected on the same local machine. These are per-call times, rather than total training times or a comparison with online planners' per-action budgets. NUTS time increases with dataset size, while SARSOP time depends strongly on the sampled model and can be dominated by long solves.

Table 9: Inference and planning time on RiverSwim and Random-S10A4: wall-clock seconds per call, mean $\pm$ 95% CI. Here K denotes the number of trajectories available for inference.

Environment	Component	$K = 50$	$K = 200$	$K = 500$
RiverSwim	NUTS	14.91 ± 6.55	48.63 ± 8.66	119.18 ± 35.08
	SARSOP	2.39 ± 4.61	0.02 ± 0.00	2.46 ± 4.78
Random-S10A4	NUTS	19.93 ± 0.68	124.17 ± 5.75	182.53 ± 5.18
	SARSOP	56.49 ± 16.45	105.59 ± 9.82	82.04 ± 15.02

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: The abstract and introduction accurately reflect the paper's contributions or scope. All main claims are supported by theoretical results in Section 4 (and proofs in appendices) and/or experimental results in Section 6 (and more experimental results in appendices).

   Guidelines:

   - The answer [N/A] 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 [N/A] 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 discuss limitations in Section 7.

   Guidelines:

   - The answer [N/A] 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]

Justification: We clearly state and label all assumptions in Section 3. Complete proofs are provided in Appendices C–D (regret upper bounds).

Guidelines:

- The answer [N/A] 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: We report all relevant information to reproduce experimental results in Section 6 and Appendices G.

   Guidelines:

   - The answer [N/A] 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: Code is available at `https://github.com/dongzeye/ps4pomdp`.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.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://neurips.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) necessary to understand the results?

Answer: [Yes]

Justification: We report all relevant information to reproduce and understand the experimental results in Section 6 and Appendices G.2, G.3, G.4, G.5, G.1, and G.2.

Guidelines:

- The answer [N/A] 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: All experimental results are reported with 95% confidence intervals over independent runs (Figure 1). Per-environment regret curves with confidence intervals are reported in Appendices G.2, G.3, G.4, and G.5.

Guidelines:

- The answer [N/A] 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).
- 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: All experiments use minimal compute. PS4POMDP runs on a single CPU per seed (no GPU required); each seed of the PPO baselines runs on a single GPU. No cluster-scale resources are needed to reproduce the results.

   Guidelines:

   - The answer [N/A] 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: The research presented in this paper conform, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] 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: [No]

    Justification: We believe there is no negative societal impact that our work might lead to directly. This work is about basic research, while certain positive societal impact may arise from it (e.g., medical applications that are inherently partially observable), none of which we believe should be discussed specifically.

    Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] 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., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] 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: [Yes]

    Justification: The creators or original owners of assets used in the paper are properly credited and license information fully disclosed.

    Guidelines:

    - The answer [N/A] 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.
- 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: [Yes]

    Justification: The paper release a few new assets (code), such as an implementation of the proposed algorithm along with a new synthetic POMDPs environments

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: the paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] 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.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.