Online Learning for Unknown Partially Observable MDPs

Mehdi Jafarnia-Jahromi
Uber Technologies, Inc.

Rahul Jain
University of Southern California
Center for Autonomy and AI

Ashutosh Nayyar
University of Southern California
Center for Autonomy and AI

Abstract

Solving Partially Observable Markov Decision Processes (POMDPs) is hard. Learning optimal controllers for POMDPs when the model is unknown is harder. Online learning of optimal controllers for unknown POMDPs, which requires efficient learning using regret-minimizing algorithms that effectively tradeoff exploration and exploitation, is even harder, and no solution exists currently. In this paper, we consider infinite-horizon average-cost POMDPs with unknown transition model, though a known observation model. We propose a natural posterior sampling-based reinforcement learning algorithm (PSRL-POMDP) and show that it achieves a regret bound of $O(\log T)$, where T is the time horizon, when the parameter set is finite. In the general case (continuous parameter set), we show that the algorithm achieves $\tilde{O}(T^{2=3})$ regret under two technical assumptions. To the best of our knowledge, this is the first online RL algorithm for POMDPs and has sub-linear regret.

1 Introduction

Reinforcement learning (RL) considers the sequential decision-making problem of an agent in an unknown environment with the goal of minimizing the total cost. The agent faces a fundamental exploration-exploitation trade-off: should it exploit the available information to minimize the cost or should it explore the environment to gather more information for future decisions? Maintaining a proper balance between exploration and exploitation is a fundamental challenge in RL and is measured with the notion of cumulative regret: the difference between the cumulative cost of the learning algorithm and that of the best policy.

The problem of balancing exploration and exploitation in RL has been successfully addressed for MDPs and algorithms with near optimal regret bounds known [Bartlett and Tewari, 2009, Jaksch et al., 2010, Ouyang et al., 2017b, Azar et al., 2017, Fruit et al., 2018, Jin et al., 2018, Abbasi-Yadkori et al., 2019b, Zhang and Ji, 2019, Zanette and Brunskill, 2019, Hao et al., 2020, Wei et al., 2020, 2021]. MDPs assume that the state is perfectly observable by the agent and the only uncertainty is about the underlying dynamics of the environment. However, in many real-world scenarios such as robotics, healthcare and finance, the state is not fully observed by the agent, and only a partial observation is available. These scenarios are modeled by Partially Observable Markov Decision Processes (POMDPs). In addition to the uncertainty in the environment dynamics, the agent has to deal with the uncertainty about the underlying state. It is well known [Kumar and Varaiya, 2015] that introducing an information or belief state (a posterior distribution over the states given the history of observations and actions) allows the POMDP to be recast as an MDP over the belief state space. The resulting algorithm requires a posterior update of the belief state which needs the transition and observation model to be fully known. This presents a significant difficulty when the model parameters are unknown. Thus, managing the exploration-exploitation trade-off for POMDPs is a significant challenge and to the best of our knowledge, no online RL algorithm with sub-linear regret is known.

In this paper, we consider infinite-horizon average-cost POMDPs with finite states, actions and observations. The underlying state transition dynamics is unknown, though we assume the observation kernel to be known. We propose a Posterior Sampling Reinforcement Learning algorithm (PSRL-POMDP) and prove that it achieves a Bayesian expected regret bound of $O(\log T)$ in the finite (transition kernel) parameter set case where T is the time horizon. We then show that in the general (continuous parameter set) case, it achieves $\tilde{O}(T^{2=3})$

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

under some technical assumptions. The PSRL-POMDP algorithm is a natural extension of the TSDE algorithm for MDPs [Ouyang et al., 2017b] with two main differences. First, in addition to the posterior distribution on the environment dynamics, the algorithm maintains a posterior distribution on the underlying state. Second, since the state is not fully observable, the agent cannot keep track of the number of visits to state-action pairs, a quantity that is crucial in the design of algorithms for tabular MDPs. Instead, we introduce a notion of pseudo count and carefully handle its relation with the true counts to obtain sub-linear regret. To the best of our knowledge, PSRL-POMDP is the first online RL algorithm for POMDPs with sub-linear regret.

1.1 Related Literature

We review the related literature in two domains: efficient exploration for MDPs, and learning in POMDPs.

Efficient exploration in MDPs. To balance the exploration and exploitation, two general techniques are used in the basic tabular MDPs: optimism in the face of uncertainty (OFU), and posterior sampling. Under the OFU technique, the agent constructs a confidence set around the system parameters, selects an optimistic parameter associated with the minimum cost from the confidence set, and takes actions with respect to the op-timistic parameter. This principle is widely used in the literature to achieve optimal regret bounds [Bartlett and Tewari, 2009, Jaksch et al., 2010, Azar et al., 2017, Fruit et al., 2018, Jin et al., 2018, Zhang and Ji, 2019, Zanette and Brunskill, 2019, Wei et al., 2020, Chen et al., 2021]. An alternative technique to encourage exploration is posterior sampling [Thompson, 1933]. In this approach, the agent maintains a posterior distribu-tion over the system parameters, samples a parameter from the posterior distribution, and takes action with respect to the sampled parameter [Strens, 2000, Os-band et al., 2013, Fonteneau et al., 2013, Gopalan and Mannor, 2015, Ouyang et al., 2017b, Jafarnia-Jahromi et al., 2021a,b]. In particular, [Ouyang et al., 2017b] proposes TSDE, a posterior sampling-based algorithm for the infinite-horizon average-cost MDPs.

Extending these results to the continuous state MDPs has been recently addressed with general function approximation [Osband and Van Roy, 2014, Dong et al., 2020, Ayoub et al., 2020, Wang et al., 2020], or in the special cases of linear function approximation [Abbasi-Yadkori et al., 2019a,b, Jin et al., 2020, Hao et al., 2020, Wei et al., 2021, Wang et al., 2021], and Linear Quadratic Regulators [Ouyang et al., 2017a, Dean et al., 2018, Cohen et al., 2019, Mania et al., 2019, Simchowitz and Foster, 2020, Lale et al., 2020a]. In general, POMDPs can be formulated as continuous state MDPs by considering the belief as the state. However, computing the belief requires the knowledge of the model parameters and thus unobserved in the RL setting. Hence, learning algorithms for continuous state MDPs cannot be directly applied to POMDPs.

Learning in POMDPs. To the best of our knowledge, the only existing work with regret analysis in POMDPs is Azizzadenesheli et al. [2017]. However, their definition of regret is not with respect to the optimal policy, but with respect to the best memoryless policy (a policy that maps the current observation to an action). With our natural definition of regret, their algorithm suffers linear regret. Other learning algorithms for POMDPs either consider linear dynamics [Lale et al., 2020b, Tsiamis and Pappas, 2020] or do not consider regret [Shani et al., 2005, Ross et al., 2007, Poupart and Vlassis, 2008, Cai et al., 2009, Liu et al., 2011, 2013, Doshi-Velez et al., 2013, Katt et al., 2018, Azizzadenesheli et al., 2018] and are not directly comparable to our setting.

Subsequent to our work, Xiong et al. [2021] also proved a regret bound of $\tilde{O}(T^{2=3})$ in the infinite-horizon average-cost POMDPs with an OFU-type algorithm. Their approach is based on spectral method of moments estimations for hidden Markov models and uses a different set of assumptions.

2 Preliminaries

An infinite-horizon average-cost Partially Observable Markov Decision Process (POMDP) can be specified by $(S; A; ; C; O;)$ where S is the state space, A is the action space, $C : S \quad A \; ! \; [0; 1]$ is the cost function, and O is the set of observations. Here $: S \; ! \; _O$ is the observation kernel, and $: S \quad A \; ! \; _S$ is the transition kernel such that $(ojs) = P(o_t = ojs_t = s)$ and $(s^0js; a) = P(s_{t+1} = s^0js_t = s; a_t = a)$ where $o_t \; 2 \; O$, $s_t \; 2 \; S$ and $a_t \; 2 \; A$ are the observation, state and action at time $t = 1; 2; 3; .$ Here, for a finite set X, $_X$ is the set of all probability distributions on X. We assume that the state space, the action space and the observations are finite with size $jSj; jAj; jOj$, respectively.

Let F_t be the information available at time t (prior to action a_t), i.e., the sigma algebra generated by the history of actions and observations $a_1; o_1; \; ; a_{t \; 1}; o_{t \; 1}; o_t$ and let F_{t+} be the information after choosing action a_t. Unlike MDPs, the state is not observable by the agent and the optimal policy cannot be a function of the state. Instead, the agent maintains a belief $h_t(;) \; 2 \; _S$ given by $h_t(s;) := P(s_t = sjF_t;)$, as a sufficient statistic for the history of observations and actions. Here we use the notation $h_t(;)$ to explicitly show the dependency of the belief on . After taking action a_t and observing

o_{t+1}, the belief h_t can be updated as $h_{t+1}(s';) = \frac{\sum_s (o_{t+1}|s')(s'|s;a_t)h_t(s;)}{\sum_{s'}\sum_s (o_{t+1}|s')(s'|s;a_t)h_t(s;)}$: (1)

This update rule is compactly denoted by $h_{t+1}(;) = (h_t(;);a_t;o_{t+1};)$, with the initial condition

$$h_1(s;) = \frac{(o_1|s)h(s)}{\sum_s (o_1|s)h(s)};$$

where $h()$ is the distribution of the initial state s_1 (denoted by s_1 h). A deterministic stationary policy $: S \to A$ maps a belief to an action. The long-term average cost of a policy $$ can be defined as

$$J(h;) := \limsup_{T\to 1} E\left[\frac{1}{T}\sum_{t=1}^{T} C(s_t;a_t) \;\middle|\; h, h_t(;) = ;\right] \quad (2)$$

Let $J(h;) := \inf J(h;)$ be the optimal long-term average cost that in general may depend on the initial state distribution h, though we will assume it is independent of the initial distribution h (and thus denoted by $J()$), and the following Bellman equation holds:

Assumption 1 (Bellman optimality equation). There exist $J() \in \mathbb{R}$ and a bounded function $v(;) : S \to \mathbb{R}$ such that for all $b \in \Delta_S$, $J() + v(b;) =$

$$\min_{a\in A}\{c(b;a) + \sum_{o\in O} P(o|b;a;)v(b';)\}; \quad (3)$$

where v is called the relative value function, $b' = (b;a;o;)$ is the updated belief, $c(b;a) := \sum_s C(s;a)b(s)$ is the expected cost, and $P(o|b;a;)$ is the probability of observing o in the next step, conditioned on the current belief b and action a, i.e.,

$$P(o|b;a;) = \sum_{s'\in S}\sum_{s\in S} (o|s')(s'|s;a)b(s): \quad (4)$$

Various conditions are known under which Assumption 1 holds, e.g., when all the entries of the transition and observation kernels are positive [Xiong et al., 2021], or when the MDP is weakly communicating [Bertsekas, 2017]. Note that if Assumption 1 holds, the policy that minimizes the right hand side of (3) is the optimal policy. More precisely,

Lemma 1. Suppose Assumption 1 holds. Then, the policy $(;) : S \to A$ given by $(b;) :=$

$$\arg\min_{a\in A}\{c(b;a) + \sum_{o\in O} P(o|b;a;)v(b';)\} \quad (5)$$

is the optimal policy with $J(h;) = J(); \forall h \in \Delta_S$.

Note that if v satisfies the Bellman equation, so does v plus any constant. Therefore, without loss of generality, and since v is bounded, we can assume that $\inf_{b\in\Delta_S} v(b;) = 0$ and define the span of a POMDP as $sp() := \sup_{b\in\Delta_S} v(b;)$. Let $_H$ be the class of POMDPs that satisfy Assumption 1 and have $sp() \le H$ for all $\in {}_H$. In Section 4, we consider a finite subset $_H$ of POMDPs. In Section 5, the general class $= {}_H$ is considered.

The learning protocol. We consider the problem of an agent interacting with an unknown randomly generated POMDP , where $\in$ is randomly generated according to the probability distribution $f()$.[1] After the initial generation of , it remains fixed, but unknown to the agent. The agent interacts with the POMDP in T steps. Initially, the agent starts from state s_1 that is randomly generated according to the conditional probability mass function $h(;)$. At time $t = 1;2;3;\ ;T$, the agent observes o_t $(|s_t)$, takes action a_t and suffers cost of $C(s_t;a_t)$. The environment, then determines the next state s_{t+1} which is randomly drawn from the probability distribution $(|s_t;a_t)$. Note that although the cost function C is assumed to be known, the agent cannot observe the value of $C(s_t;a_t)$ since the state s_t is unknown to the agent. The goal of the agent is to minimize the expected cumulative regret defined as

$$R_T := E\left[\sum_{t=1}^{T} C(s_t;a_t) - J() \;\middle|\; ;\right] \quad (6)$$

where the expectation is with respect to the prior distribution $h(;)$ for s_1, the randomness in the state transitions, and the randomness in the algorithm. Here, $E[]$ is a shorthand for $E[|]$. In Section 4, a regret bound is provided on R_T, however, Section 5 considers $E[R_T]$ (also called Bayesian regret) as the performance measure for the learning algorithm. We note that the Bayesian regret is widely considered in the MDP literature [Osband et al., 2013, Gopalan and Mannor, 2015, Ouyang et al., 2017b,a].

3 The PSRL-POMDP Algorithm

We propose a general Posterior Sampling Reinforcement Learning for POMDPs (PSRL-POMDP) algorithm (Algorithm 1) for both the finite-parameter and the general case. The algorithm maintains a joint distribution on the unknown parameter as well as the state s_t. PSRL-POMDP takes the prior distributions h and f as input. At time t, the agent computes the posterior distribution $f_t()$ on the unknown parameter as well as the posterior conditional probability mass function (pmf) $h_t(;)$ on the state s_t for $\in$. Upon taking

[1] In Section 4, $f()$ should be viewed as a probability mass function.

action a_t and observing o_{t+1}, the posterior distribution is updated by applying the Bayes' rule as[2]

$$f_{t+1}(\theta) = \frac{\sum_{s,s'} O_\theta(o_{t+1}|s')P_\theta(s'|s,a_t)h_t(s,\theta)f_t(\theta)}{\int_\Theta \sum_{s,s'} O_\theta(o_{t+1}|s')P_\theta(s'|s,a_t)h_t(s,\theta)f_t(\theta)d\theta};$$

$$h_{t+1}(s',\theta) = \tau(h_t(\cdot,\theta),a_t,o_{t+1},s',\theta), \quad (7)$$

with the initial condition

$$f_1(\theta) = \frac{\sum_s O_\theta(o_1|s)h(s,\theta)f(\theta)}{\int_\Theta \sum_s O_\theta(o_1|s)h(s,\theta)f(\theta)d\theta};$$

$$h_1(s,\theta) = \frac{O_\theta(o_1|s)h(s,\theta)}{\sum_s O_\theta(o_1|s)h(s,\theta)}. \quad (8)$$

Recall that $\tau(h_t(\cdot,\theta),a_t,o_{t+1},\theta)$ is a compact notation for (1). In the special case of perfect observation at time t, $h_t(s,\theta) = 1(s_t = s)$ for all $\theta \in \Theta$ and $s \in \mathcal{S}$. Moreover, the update rule of f_{t+1} reduces to that of fully observable MDPs (see Eq. (4) of Ouyang et al. [2017b]) in the special case of perfect observation at time t and $t+1$.

Let $n_t(s,a) = \sum_{\tau=1}^{t-1} 1(s_\tau = s, a_\tau = a)$ be the number of visits to state-action (s,a) by time t. The num-ber of visits n_t plays an important role in learning for MDPs [Jaksch et al., 2010, Ouyang et al., 2017b] and is one of the two criteria to determine the length of the episodes in the TSDE algorithm for MDPs [Ouyang et al., 2017b]. However, in POMDPs, n_t is not $\mathcal{F}_{(t-1)+}$-measurable since the states are not observable. Instead, let $\bar{n}_t(s,a) := \mathbb{E}[n_t(s,a)|\mathcal{F}_{(t-1)+}]$, and define the pseudo-count m_t as follows:

Definition 1. $(m_t)_{t=1}^T$ is a pseudo-count if it is a non-decreasing, integer-valued sequence of random variables such that m_t is $\mathcal{F}_{(t-1)+}$-measurable, $m_t(s,a) \geq \lceil \bar{n}_t(s,a) \rceil$, and $m_t(s,a) \geq t$ for all $t \leq T+1$.

An example of such a sequence is simply $m_t(s,a) = t$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. This is used in Section 4. Another example is $m_t(s,a) := \max\{m_{t-1}(s,a), \lceil \bar{n}_t(s,a) \rceil\}$ with $m_0(s,a) = 0$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ which is used in Section 5. Here $\lceil \bar{n}_t(s,a) \rceil$ is the smallest integer that is greater than or equal to $\bar{n}_t(s,a)$. By definition, m_t is integer-valued and non-decreasing which is essential to bound the number of episodes in the algorithm for the general case (see Lemma B.5).

Similar to the TSDE algorithm for fully observable MDPs, PSRL-POMDP algorithm proceeds in episodes. In the beginning of episode k, POMDP θ_k is sampled from the posterior distribution f_{t_k} where t_k denotes the start time of episode k. The optimal policy $\mu(\cdot,\theta_k)$ is then computed and used during the episode. Note that the input of the policy is $h_t(\cdot,\theta_k)$. The intuition behind such a choice (as opposed to the belief $b_t(\theta) := \int_\Theta h_t(\cdot,\theta)f_t(\theta)d\theta$) is that during episode k, the agent treats θ_k to be the true POMDP and adopts the optimal policy with respect to it. Consequently, the input to the policy should also be the conditional belief with respect to the sampled θ_k.

[2] When the parameter set Θ is finite, $\int_\Theta$ should be replaced with $\sum_\Theta$.

Algorithm 1: PSRL-POMDP

Require: prior distributions $f(\cdot)$; $h(\cdot)$
  Initialization: $t \leftarrow 1$; $t_1 \leftarrow 0$
  Observe o_1 and compute f_1; h_1 according to (8)
1: **for** episodes $k = 1, 2, \ldots$ **do**
2: $T_{k-1} \leftarrow t - t_k$
3: $t_k \leftarrow t$
4: Generate $\theta_k \sim f_t(\cdot)$ and compute $\mu_k(\cdot) = \mu(\cdot,\theta_k)$ from (5)
5: **while** $t < \mathrm{SCHED}(t_k, T_{k-1})$ and $m_t(s,a) < 2m_{t_k}(s,a)$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ **do**
6: Choose action $a_t = \mu_k(h_t(\cdot,\theta_k))$ and observe o_{t+1}
7: Update f_{t+1}; h_{t+1} according to (7)
8: $t \leftarrow t+1$
9: **end while**
10: **end for**

A key factor in designing posterior sampling based algorithms is the design of episodes. Let T_k denote the length of episode k. In PSRL-POMDP, a new episode starts if either $t > \mathrm{SCHED}(t_k, T_{k-1})$ or $m_t(s,a) > 2m_{t_k}(s,a)$. In the finite parameter case (Section 4), we consider $\mathrm{SCHED}(t_k, T_{k-1}) = 2t_k$ and $m_t(s,a) = t$. With these choices, the two crite-ria coincide and ensure that the start time and the length of the episodes are deterministic. In Sec-tion 5, we use $\mathrm{SCHED}(t_k, T_{k-1}) = t_k + T_{k-1}$ and $m_t(s,a) := \max\{m_{t-1}(s,a), \lceil \bar{n}_t(s,a) \rceil\}$. This guaran-tees that $T_k \leq T_{k-1} + 1$ and $m_t(s,a) \leq 2m_{t_k}(s,a)$. These criteria are previously introduced in the TSDE algorithm [Ouyang et al., 2017b] except that TSDE uses the true count n_t rather than m_t.

4 Finite-Parameter Case ($|\Theta| < \infty$)

In this section, we consider Θ such that $|\Theta| < \infty$. When Θ is finite, the posterior distribution concentrates on the true parameter exponentially fast if the tran-sition kernels are separated enough (see Lemma 2). This allows us to achieve a regret bound of $O(H \log T)$. Let $o_{1:t}, a_{1:t}$ be a shorthand for the history of observations $o_1, \ldots, o_t$ and actions $a_1, \ldots, a_t$, respectively. Let $P_\theta^{o_{1:t},a_{1:t}}(o)$ be the probability of observing o at time $t+1$ if the action history is $a_{1:t}$, the observation history is $o_{1:t}$, and the transition kernel is θ, i.e.,

$$P_\theta^{o_{1:t},a_{1:t}}(o) := P(o_{t+1} = o | o_{1:t}, a_{1:t}, \theta = \theta).$$

The distance between $o_{1:t};a_{1:t}$ and $o_{1:t};a_{1:t}$ is defined by Kullback Leibler (KL-) divergence as follows. For a fixed state-action pair $(s;a)$ and any $;\ 2$, denote by $K(^{o_{1:t};a_{1:t}}k^{o_{1:t};a_{1:t}})$, the Kullback Leibler (KL-) divergence between the probability distributions $o_{1:t};a_{1:t}$ and $o_{1:t};a_{1:t}$ is given by

$$K(^{1:o\ 1;a}k^{1:t\ o:t});a = \sum_{o}{}^{1:t\ 1:t}(o)\log {}_{o_{1:t};a_{1:t}}\frac{o_{1:t};a_{1:t}(o)}{\cdot}$$

It can be shown that $K(^{o_{1:t};a_{1:t}}k^{o_{1:t};a_{1:t}})$ 0 and that equality holds if and only if $^{o_{1:t};a_{1:t}} = {}_{o_{1:t};a_{1:t}}$. Thus, KL-divergence can be thought of as a measure of divergence of $^{o_{1:t};a_{1:t}}$ from $o_{1:t};a_{1:t}$. In this section, we need to assume that the transition kernels in are distant enough in the following sense.

Assumption 2. There exist positive constants > 0 and $B > 0$ such that for any time step t, any history of possible observations $o_{1:t}$ and actions $a_{1:t}$, and any two transition kernels $;\ 2$ such that $^{o_{1:t\ 1};a_{1:t\ 1}}(o\ )_t > 0$, we have $K(^{o_{1:t};a_{1:t}}k^{o_{1:t};a_{1:t}})$ and $^{o_{1:t\ 1};a_{1:t\ 1}}(o_t)/{}^{o_{1:t\ 1};a_{1:t\ 1}}(o_t)$ B.

This assumption is similar to that of Kim [2017].

Theorem 1. Suppose Assumptions 1 and 2 hold. Then, the regret bound of Algorithm 1 with $\mathrm{SCHED}(t_k;T_{k\ 1}) = 2t_k$ and $\hat{m}_t(s;a) = t$ for all state-action pairs $(s;a)$ is bounded as

$$R_T \quad H\log T + \frac{4(H+1)}{(\ e\qquad 1)^2}.$$

where > 0 is a universal constant defined in Lemma 2.

Observe that with $\mathrm{SCHED}(t_k;T_{k\ 1}) = 2t_k$ and $\hat{m}_t(s;a) = t$, the two stopping criteria in Algorithm 1 coincide and ensure that $T_k = 2T_{k\ 1}$ with $T_0 = 1$. In other words, the length of episodes grows as $T_k = 2^k$.

4.1 Proof of Theorem 1

In this section, proof of Theorem 1 is provided. A key factor in achieving $O(H\log T)$ regret bound in the case of finite parameters is that the posterior distribution $f_t()$ concentrates on the true exponentially fast.

Lemma 2. Suppose Assumption 2 holds. Then, there exist constants > 1 and > 0 such that $E[1\ f_t()j]\ \exp(\ t)$.

Proof. Let $_t = fa_1;o_1;\ ;a_{t\ 1};o_{t\ 1};o_tg$ be the trajectory of actions and observations and define the likelihood function

$$L(_tj) := P(_tj) = P(o_1j)\prod_{=2}^{t} P(o_jo_{1:\ 1};a_{1:\ 1};)$$

$$= P(o_1j)\prod_{=2}^{t}{}^{o_{1:\ 1};a_{1:\ 1}}(o)$$

Note that $P(o_1j) = \sum_s h(s)(o_1js)$ is independent of , thus for any $;\ 2$ such that $L(_tj) = 0$ and $L(_tj) = 0$, we can write

$$\frac{L(_tj)}{} = \prod_{=2}^{t}\frac{{}^{o_{1:\ 1};a_{1:\ 1}}(o)\ L(_tj)}{{}^{o_{1:\ 1};a_{1:\ 1}}(o)}:$$

Recall that $f_t()$ is the posterior associated with the likelihood given by

$$f_t() = \frac{L(_tj)f()}{\sum_{2}L(_tj)f()}:$$

In the denominator, we exclude those such that $L(_tj) = 0$ without loss of generality. We now proceed to lower bound $f_t()$ for those such that $L(_tj) > 0$. We can write

$$f_t() = \frac{L(_tj)f()}{\sum L(_tj)f()} = \frac{1}{1+\sum_{=}\frac{f()}{f()}\frac{L(\ j)}{L(_tj)}}$$

$$= \frac{1}{1+\sum_{=}f()\exp(\ \sum_{=1}^{t}\log\ ;)},$$

where we define ${}_1 := 1$ and for 2,

$$; := \frac{{}^{o_{1:\ 1};a_{1:\ 1}}(o)}{o\quad ;a_{\ 1:\ 1}(o)}:$$

Denote by $Z_t^{;} := \sum_{=1}^{t}\log\ ;$ and decompose it as $Z^{;} = {}_tM^{;} + A_t^{;}$ where

$$M_t^{;} := \sum^{t}\log\ ;\quad E[\log\ ;F_{\ 1};\ =\ ];\ =1$$

$$A_t^{;} := \sum^{t}E[\log\ ;F_{\ 1};\ =\ :=1\ ]$$

Note that the terms inside the first summation constitute a martingale difference sequence with respect to the filtration $(F)_1$ and conditional probability $P(j\ =\ )$. Each term is bounded as $j\log\ ;$ $E[\log\ ;jF_{\ 1};\ =\ ]j\ d$ for some $d > 0$ by Assumption 2. The second term, $A_t^{;}$ can be lower bounded using Assumption 2 as follows

$$E[\log\ ;F_{\ 1};\ =$$

$$= E[E[\log\ ;F_{\ 1};a_{\ 1};\ =\ F_{\ 1};\ =]$$

$$= E[K(^{o_{1:\ 1};a_{1:\ 1}}k^{o_{1:\ 1};a_{1:\ 1}})F_{\ 1};\ =$$

Summing over implies that

$$A_t^{;}\ t: \tag{9}$$

To bound M'_t, let $0 < \epsilon < \gamma$, and apply Azuma's inequality to obtain $P(|M'_t| \geq t) \leq 2\exp(-\frac{t^2}{2d^2})$.

Fix ϵ. Union bound over all $\gamma = \epsilon$ implies that the event $B'_t := \bigcap_{\gamma=\epsilon} \{|M'_t| < t\}$ happens with probability at least $1 - 2(|\mathcal{J}| - 1)\exp(-\frac{t^2}{2d^2})$. If B'_t holds, then $M'_t < t$ for all $\gamma = \epsilon$. Combining this with (9) implies that $\exp(M'_t - A'_t) \leq \exp(t - t)$. Therefore, $\mathbb{E}[f_t(\theta) \mid \gamma = \epsilon] \leq$

$$\mathbb{E}\left[\frac{1}{1 + \sum_{\theta' \neq \theta} \frac{f(\theta')}{f(\theta)} \exp(M'_t - A'_t)}\right] \leq \mathbb{E}\left[\frac{\mathbb{1}(B'_t)}{1 + \sum_{\theta' \neq \theta} \frac{f(\theta')}{f(\theta)} \exp(t - t)}\right] + P(B'^t)$$

$$\leq \frac{1}{1 + \frac{f(\theta')}{f(\theta)} \exp(t - t)} + 2(|\mathcal{J}| - 1)\exp\left(-\frac{t^2}{2d^2}\right).$$

Now, by choosing $\epsilon = \gamma/2$, and constants $t \geq 2\max\{\max_{\theta' \neq \theta} \frac{2}{1 - \frac{f(\theta')}{f(\theta)}}, 2(|\mathcal{J}| - 1)\}$, $t \geq \min\{t : t \geq 8d^2 \log t\}$, we have

$$\mathbb{E}[1 - f_t(\theta) \mid \gamma = \epsilon] \leq 1 - \frac{1 - 2(|\mathcal{J}| - 1)\exp(-\frac{t^2}{2d^2})}{1 + \frac{1 - f(\theta)}{f(\theta)} \exp(t - t)}$$

$$= \frac{\frac{1 - f(\theta)}{f(\theta)} \exp(t - t) + 2(|\mathcal{J}| - 1)\exp(-\frac{t^2}{2d^2}) \cdot 1 + \frac{2}{2}}{\frac{f(\theta)}{f(\theta)} \exp(t - t)}$$

$$\leq \frac{1}{f(\theta)} \exp(t - t) + 2(|\mathcal{J}| - 1)\exp\left(-\frac{t^2}{2d^2}\right) = \frac{1}{2}$$

$$f(\theta) \exp\left(-\frac{t}{f(\theta)}\right) + 2(|\mathcal{J}| - 1)\exp\left(-\frac{t^2}{8d^2}\right) \leq \exp(-t).$$

Equipped with this lemma, we are now ready to prove Theorem 1. □

Proof. Let K_T be the number of episodes by time T. Note that the regret R_T can be decomposed as $R_T = H\,\mathbb{E}[K_T] + R_1 + R_2 + R_3$ by Lemma A.1, where

$$R_1 := \mathbb{E}\left[\sum_{k=1}^{K_T} T_k \left[J(\theta_k) - J(\theta)\right]\right];$$

$$R_2 := H\,\mathbb{E}\left[\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \sum_{s'} |\mathcal{P}(s'|s_t, a_t) - \mathcal{P}_{\theta_k}(s'|s_t, a_t)|\right]$$

$$+ \sum_{s} |h_t(s, \theta) - h_t(s, \theta_k)|\Bigg];$$

$$R_3 := \mathbb{E}\left[\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \left[c(h_t(\cdot, \theta), a_t) - c(h_t(\cdot, \theta_k), a_t)\right]\right].$$

Note that the start time and length of episodes in Algorithm 1 are deterministic with the choice of SCHED and m_t in the statement of the theorem, i.e., t_k, T_k and hence K_T are deterministic. Note that if $\theta_k = \theta$, then $R_1 = R_2 = R_3 = 0$. Moreover, we have that $|J(\theta_k) - J(\theta)| \leq 1$, $\sum_{s'} |\mathcal{P}(s'|s_t, a_t) - \mathcal{P}_{\theta_k}(s'|s_t, a_t)| \leq 2$, $\sum_{s} |h_t(s, \theta) - h_t(s, \theta_k)| \leq 2$, and $|c(h_t(\cdot, \theta), a_t) - c(h_t(\cdot, \theta_k), a_t)| \leq 1$. Therefore,

$$R_1 := \mathbb{E}\left[\sum_{k=1}^{K_T} T_k \mathbb{1}(\theta_k = \theta)\right] = \sum_{k=1}^{K_T} T_k P(\theta_k = \theta); \quad \theta_{k=1}$$

$$R_2 := 4H\,\mathbb{E}\left[\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \mathbb{1}(\theta_k = \theta)\right]$$

$$= 4H \sum_{k=1}^{K_T} T_k P(\theta_k = \theta); \quad \theta_{k=1}$$

$$R_3 := \mathbb{E}\left[\sum_{t=t_k}^{t_{k+1}-1} \sum_{k=1}^{K_T} \mathbb{1}(\theta_k = \theta)\right] = \sum_{k=1}^{K_T} T_k P(\theta_k = \theta). \quad \theta_{k=1}$$

Note that $P(\theta_k = \theta) = \mathbb{E}[1 - f_t(\theta)] \leq \exp(-t_k)$ by Lemma 2. Combining all these bounds, we can write

$$R_T \leq H K_T + (4H + 2) \sum_{k=1}^{K_T} T_k \exp(-t_k).$$

With the episode schedule provided in the statement of the theorem, it is easy to check that $K_T = O(\log T)$. Let $n = 2^{K_T}$ and write

$$\sum_{k=1}^{K_T} T_k \exp(-t_k) = \sum_{k=1}^{K_T} 2^k e^{-(2^k - 1)}$$

$$\sum_{j=2}^{n} j e^{-(j-1)} = \frac{d}{dx} \frac{x^{n+1} - 1}{x - 1}\Bigg|_{x=e} - 1.$$

The last equality is by geometric series. Simplifying the derivative yields

$$\frac{d}{dx} \frac{x^{n+1} - 1}{x - 1}\Bigg|_{x=e} = \frac{nx^{n+1} - (n+1)x^n + 1}{(x-1)^2}\Bigg|_{x=e}$$

$$\leq \frac{x^n + 1}{(x-1)^2}\Bigg|_{x=e} \leq \frac{2}{(e-1)^2}.$$

Substituting these values implies $R_T \leq H \log T + \frac{4(H+1)}{(e-1)^2}$. □

5 General Case ($|\mathcal{J}| = 1$)

We now consider the general case, where the parameter set is infinite, and in particular, $\Theta = \mathbb{R}^H$, an uncountable set. We make the following two technical assumptions on the belief and the transition kernel.

Assumption 3. Denote by k(t) the episode at time t. The true conditional belief $h_t(\cdot)$ and the approximate conditional belief $h_t(\cdot;\theta_{k(t)})$ satisfy

$$\mathbb{E}\Big[\sum_s h_t(s;\theta) \quad h_t(s;\theta_{k(t)})\Big] \quad K_1(|S||A||O|)\sqrt{\frac{\log t}{t}}; \theta_{k(t)} \tag{10}$$

with probability at least 1 , for any $2(0;1)$. Here $K_1(|S||A||O|)$ is a constant that is polynomial in its input parameters and hides the logarithmic dependency on $|S||A||O|;T;$.

Assumption 3 states that the gap between conditional posterior function for the sampled POMDP $_k$ and the true POMDP decreases with episodes as better approximation of the true POMDP is available. There has been recent work on computation of approximate information states as required in Assumption 3 [Subramanian et al., 2020].

Assumption 4. There exists an F_t-measurable estimator $\hat{}_t : S \times A \to \Delta_S$ such that

$$\sum_{s'} |(s'|s;a) \quad \hat{}_t(s'|s;a)| \quad K_2(|S||A||O|)\sqrt{\frac{|S|}{\max\{1;\tilde{n}_t(s;a)\}}} \tag{11}$$

with probability at least 1 , for any $2(0;1)$, uniformly for all $t = 1;2;3;\ldots;T$, where $K_2(|S||A||O|)$ is a constant that is polynomial in its input parameters and hides the logarithmic depen-dency on $|S||A||O|;T;$.

There has been extensive work on estimation of transition dynamics of MDPs, e.g., [Grunewalder et al., 2012]. Two examples where Assumptions 3, 4 hold are:

- Perfect observation. In the case of perfect observation, where $h_t(s;\cdot) = 1(s_t = s)$, Assumption 3 is clearly satisfied. Moreover, with perfect observation, one can choose $\tilde{n}_t(s;a) = n_t(s;a)$ and select $\hat{}_k(s'|s;a) = \frac{n_t(s;a;s')}{n_t(s;a)}$ to satisfy Assumption 4 [Jaksch et al., 2010, Ouyang et al., 2017b]. Here $n_t(s;a;s')$ denotes the number of visits to $s;a$ such that the next state is s' before time t.
- Finite-parameter case. In the finite-parameter case with the choice of $\tilde{n}_t(s;a) = t$ for all state-action pairs $(s;a)$ and $\text{SCHED}(t_k;T_{k\ 1}) = t_k + T_{k\ 1}$ or $\text{SCHED}(t_k;T_{k=1}) = 2t_k$, both of the assumptions are satisfied (see Lemma B.1 for details). Note that in this case a more refined analysis is performed in Section 4 to achieve $O(H\log T)$ regret bound.

Now, we state the main result of this section.

Theorem 2. Under Assumptions 1, 3 and 4, running PSRL-POMDP algorithm with $\text{SCHED}(t_k;T_{k\ 1}) = t_k + T_{k\ 1}$ yields $\mathbb{E}[R_T] \quad O(\tilde{H}K_2(|S||A||O|T)^{2=3})$, where $K_2 := K_2(|S||A||O|)$ in Assumption 4.

The exact constants are known (see proof and Appendix B.1) though we have hidden them above.

5.1 Proof Sketch of Theorem 2

We provide the proof sketch of Theorem 2 here. A key property of posterior sampling is that conditioned on the information at time t, the sampled $_t$ and the true have the same distribution [Osband et al., 2013, Russo and Van Roy, 2014]. Since the episode start time t_k is a stopping time with respect to the filtration $(F_t)_{t1}$, we use a stopping time version of this property:

Lemma 3 (Lemma 2 in Ouyang et al. [2017b]). For any measurable function g and any F_{t_k}-measurable random variable X, we have $\mathbb{E}[g(_k;X)] = \mathbb{E}[g(;X)]$.

Introducing the pseudo count $\tilde{n}_t(s;a)$ in the algorithm requires a novel analysis to achieve a low regret bound. The following key lemma states that the pseudo count $\tilde{n}_t$ cannot be too smaller than the true count n_t.

Lemma 4. Fix a state-action pair $(s;a) \in S\times A$. For any pseudo count $\tilde{n}_t$ and any $2[0;1]$,

$$P\{\tilde{n}_t(s;a) < n_t(s;a)\} \quad : \tag{12}$$

Proof. We show that $P\{\tilde{n}_t(s;a) < n_t(s;a)\} \quad$. Since by definition $\tilde{n}_t(s;a) \quad n_t(s;a)$, the claim of the lemma follows. For any $2[0;1]$,

$$\tilde{n}_t(s;a)1\{\ n_t(s;a) > \tilde{n}_t(s;a)\} \quad n_t(s;a)1\{\ n_t(s;a) > \tilde{n}_t(s;a)\}. \tag{13}$$

By taking conditional expectation with respect to $F_{(t\ 1)+}$ from both sides and the fact that $\mathbb{E}[n_t(s;a)|F_{(t\ 1)+}] = \tilde{n}_t(s;a)$, we have

$$\tilde{n}_t(s;a)\mathbb{E}\big[1\{\ n_t(s;a) > \tilde{n}_t(s;a)\}\ F_{(t\ 1)+}\big] \quad \tilde{n}_t(s;a)1\{\ n_t(s;a) > \tilde{n}_t(s;a)\}. \tag{14}$$

We claim that

$$\mathbb{E}\big[1\{\ n_t(s;a) > \tilde{n}_t(s;a)\}|F_{(t\ 1)+}\ ;\big] \text{ a.s.} \tag{15}$$

If this claim is true, taking another expectation from both sides completes the proof.

To prove the claim, let $_0;\ _+$ be the subsets of the sample space where $\tilde{n}_t(s;a) = 0$ and $\tilde{n}_t(s;a) > 0$, respectively. We consider these two cases sepa-rately: (a) on $_+$ one can divide both sides of (14) by $\tilde{n}_t(s;a)$ and reach (15); (b) note that by defini-tion $\tilde{n}_t(s;a) = 0$ on $_0$. Thus, $n_t(s;a)1(_0) = 0$ almost surely (this is because $\mathbb{E}[n_t(s;a)1(_0)] = \mathbb{E}[\mathbb{E}[n_t(s;a)1(_0)|F_{(t\ 1)+}]] = \mathbb{E}[\tilde{n}_t(s;a)1(_0)] = 0$). Therefore,

$$1(_0)1\{\ n_t(s;a) > \tilde{n}_t(s;a) = 0; \quad \text{a.s.},$$

which implies

$$\left(1 - \frac{1}{\alpha_0}\right) \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left(n_t(s,a) > \hat{m}_t(s,a)\right) \mid \mathcal{F}_{(t-1)^+}\right] = 0, \quad \text{a.s.},$$

which means on α_0, the left hand side of (15) is indeed zero, almost surely, proving the claim. □

The parameter α will be tuned later to balance two terms and achieve $\tilde{O}(T^{2/3})$ regret bound (see Lemma B.4). We are now ready to provide the proof sketch of Theorem 2.

By Lemma A.1, R_T can be decomposed as $R_T = H\,\mathbb{E}[K_T] + R_1 + R_2 + R_3$, where

$$R_1 := \mathbb{E}\left[\sum_{k=1}^{K_T} \left[T_k J(\theta_k) - J(\theta^*)\right]\right];$$

$$R_2 := H\,\mathbb{E}\left[\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \sum_{s'} \left|\mathbb{P}(s'|s_t,a_t) - \mathbb{P}_{\theta_k}(s'|s_t,a_t)\right|\right] + \mathbb{E}\left[\sum_{t=1}^{T} \left|h_t(s,\cdot) - h_t(s,\theta_k)\right|\right];$$

$$R_3 := \mathbb{E}\left[\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \left[c(h_t(\cdot,\cdot),a_t) - c(h_t(\cdot,\theta_k),a_t)\right]\right].$$

It follows from the first stopping criterion that $T_k \le T_{k-1} + 1$. Using this along with the property of poste-rior sampling (Lemma 3) proves that $\mathbb{E}[R_1] \le \mathbb{E}[K_T]$ (see Lemma B.2 for details). $\mathbb{E}[R_3]$ is bounded by $K_1\,\mathbb{E}\left[\sum_{k=1}^{K_T} \frac{T_k - 1}{\sqrt{t_k - 1}}\right] + 1$ where $K_1 := K_1(|\mathcal{S}|, |\mathcal{A}|, |\mathcal{O}|, \cdot)$ is the constant in Assumption 3 (see Lemma B.3). To bound $\mathbb{E}[R_2]$, we use Assumption 3 and follow the proof steps of Lemma B.3 to conclude that

$$\mathbb{E}[R_2] \le R_2 + H K_1\,\mathbb{E}\left[\sum_{k=1}^{K_T} \frac{T_k}{\sqrt{t_k}}\right] + 1;$$

where

$$R_2 := H\,\mathbb{E}\left[\sum_{t=t_k}^{t_{k+1}-1} \sum_{s'} \left|\mathbb{P}(s'|s_t,a_t) - \mathbb{P}_{\theta_k}(s'|s_t,a_t)\right|\right].$$

R_2 is the dominating term in the final $\tilde{O}(T^{2/3})$ regret bound and can be bounded by $H + 12 H K_2 (|\mathcal{S}||\mathcal{A}|T)^{2/3}$ where $K_2 := K_2(|\mathcal{S}|, |\mathcal{A}|, |\mathcal{O}|, \cdot)$ is the constant in Assumption 4. The detailed proof can be found in Lemma B.4. However, we sketch the main steps of the proof here. By Assumption 4, one can show that

$$R_2 \le \tilde{O}\,\mathbb{E}\left[\sum_{t=1}^{T} \frac{H K_2}{\sqrt{\max\{1, \hat{m}_t(s_t,a_t)\}}}\right].$$

Now, let E_2 be the event that $\hat{m}_t(s,a) \ge n_t(s,a)$ for all s,a. Note that by Lemma 4 and union bound, $\mathbb{P}(E_2^c) \le |\mathcal{S}||\mathcal{A}|$. Thus,

$$R_2 \le \tilde{O}\,\mathbb{E}\left[\sum_{t=1}^{T} \frac{H K_2}{\sqrt{\max\{1, \hat{m}_t(s_t,a_t)\}}} \mathbb{1}(E_2) + \mathbb{1}(E_2^c)\right]$$

$$\le \tilde{O}\,\mathbb{E}\left[\sum_{t=1}^{T} \frac{K_2}{\sqrt{\max\{1, n_t(s_t,a_t)\}}}\right] + H K_2 |\mathcal{S}||\mathcal{A}| T$$

Algebraic manipulation of the inner summation yields $R_2 \le \tilde{O}\left(H K_2 \sqrt{|\mathcal{S}||\mathcal{A}|T}\right) + H K_2 |\mathcal{S}||\mathcal{A}| T$.

Optimizing over α implies $R_2 = \tilde{O}(H K_2 (|\mathcal{S}||\mathcal{A}|T)^{2/3})$. Substituting upper bounds for $\mathbb{E}[R_1], \mathbb{E}[R_2]$ and $\mathbb{E}[R_3]$, we get

$$\mathbb{E}[R_T] = H\,\mathbb{E}[K_T] + \mathbb{E}[R_1] + \mathbb{E}[R_2] + \mathbb{E}[R_3]$$

$$\le (1+H)\,\mathbb{E}[K_T] + 12 H K_2 (|\mathcal{S}||\mathcal{A}|T)^{2/3} + (H+1) K_1\,\mathbb{E}\left[\sum_{k=1}^{K_T} \frac{T_k}{\sqrt{t_k}}\right] + 2 + H.$$

From Lemma B.5, we know that $\mathbb{E}[K_T] = \tilde{O}(\sqrt{|\mathcal{S}||\mathcal{A}|T})$ and $\sum_{k=1}^{K_T} \frac{T_k}{\sqrt{t_k}} = \tilde{O}(|\mathcal{S}||\mathcal{A}|\sqrt{T})$. Therefore, $\mathbb{E}[R_T] \le \tilde{O}(H K_2 (|\mathcal{S}||\mathcal{A}|T)^{2/3})$.

Acknowledgments

MJ and RJ's research was supported by NSF awards CCF-1817212, ECCS-1810447 and ECCS-2025732. AN's research was supported by NSF awards ECCS 1750041 and ECCS 2025732, and the Okawa foundation research grant.

Conclusions

In this paper, we have presented one of the first online reinforcement learning algorithms for POMDPs. Solving POMDPs is a hard problem. Designing an efficient learning algorithm that achieves sublinear regret is even harder. We show that the proposed PSRL-POMDP algorithm achieves a Bayesian regret bound of $O(\log T)$ when the parameter is finite. When the parameter set may be uncountable, we showed a $\tilde{O}(T^{2/3})$ regret bound under two technical assumptions on the belief state approximation and transition kernel estimation. There has been recent work that does approximate belief state computation, as well as estimates transition dynamics of continuous MDPs, and in future work, we will try to incorporate such estimators. We also assume that the observation kernel is known. Note that without it, it is very challenging to design online learning algorithms for POMDPs. Posterior sampling-based algorithms in general are known to have superior numerical performance as compared to OFU-based algorithms for bandits and MDPs. In future work, we will

also do an experimental investigation of the proposed algorithm. An impediment is that available POMDP solvers mostly provide approximate solutions which would lead to linear regret. In the future, we will also try to improve the regret for the general case to $\tilde{O}(\sqrt{T})$.

References

Yasin Abbasi-Yadkori, Peter Bartlett, Kush Bhatia, Nevena Lazic, Csaba Szepesvari, and Gellért Weisz. Politex: Regret bounds for policy iteration using expert prediction. In International Conference on Machine Learning, pages 3692–3702, 2019a.

Yasin Abbasi-Yadkori, Nevena Lazic, Csaba Szepesvari, and Gellert Weisz. Exploration-enhanced politex. arXiv preprint arXiv:1908.10479, 2019b.

Alex Ayoub, Zeyu Jia, Csaba Szepesvari, Mengdi Wang, and Lin Yang. Model-based reinforcement learning with value-targeted regression. In International Conference on Machine Learning, pages 463–474. PMLR, 2020.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In Proceedings of the 34th International Conference on Machine Learning-Volume 70, pages 263–272. JMLR. org, 2017.

Kamyar Azizzadenesheli, Alessandro Lazaric, and Animashree Anandkumar. Experimental results: Reinforcement learning of pomdps using spectral methods. arXiv preprint arXiv:1705.02553, 2017.

Kamyar Azizzadenesheli, Yisong Yue, and Animashree Anandkumar. Policy gradient in partially observable environments: Approximation and convergence. arXiv e-prints, pages arXiv–1810, 2018.

Peter L Bartlett and Ambuj Tewari. Regal: A regularization based algorithm for reinforcement learning in weakly communicating mdps. In Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, pages 35–42. AUAI Press, 2009.

Dimitri P Bertsekas. Dynamic programming and optimal control, vol i and ii, 4th edition. Belmont, MA: Athena Scientific, 2017.

Chenghui Cai, Xuejun Liao, and Lawrence Carin. Learning to explore and exploit in pomdps. Advances in Neural Information Processing Systems, 22:198–206, 2009.

Liyu Chen, Mehdi Jafarnia-Jahromi, Rahul Jain, and Haipeng Luo. Implicit finite-horizon approximation and efficient optimal algorithms for stochastic shortest path. Advances in Neural Information Processing Systems, 2021.

Alon Cohen, Tomer Koren, and Yishay Mansour. Learning linear-quadratic regulators efficiently with only $\sqrt{T}$ regret. In International Conference on Machine Learning, pages 1300–1309. PMLR, 2019.

Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. Regret bounds for robust adaptive control of the linear quadratic regulator. arXiv preprint arXiv:1805.09388, 2018.

Kefan Dong, Jian Peng, Yining Wang, and Yuan Zhou. Root-n-regret for learning in markov decision processes with function approximation and low bellman rank. In Conference on Learning Theory, pages 1554–1557. PMLR, 2020.

Finale Doshi-Velez, David Pfau, Frank Wood, and Nicholas Roy. Bayesian nonparametric methods for partially-observable reinforcement learning. IEEE transactions on pattern analysis and machine intelligence, 37(2):394–407, 2013.

Raphaël Fonteneau, Nathan Korda, and Rémi Munos. An optimistic posterior sampling strategy for bayesian reinforcement learning. In NIPS 2013 Workshop on Bayesian Optimization (BayesOpt2013), 2013.

Ronan Fruit, Matteo Pirotta, Alessandro Lazaric, and Ronald Ortner. Efficient bias-span-constrained exploration-exploitation in reinforcement learning. In International Conference on Machine Learning, pages 1573–1581, 2018.

Aditya Gopalan and Shie Mannor. Thompson sampling for learning parameterized markov decision processes. In Conference on Learning Theory, pages 861–898. PMLR, 2015.

Steffen Grunewalder, Guy Lever, Luca Baldassarre, Massi Pontil, and Arthur Gretton. Modelling transition dynamics in mdps with rkhs embeddings. arXiv preprint arXiv:1206.4655, 2012.

Botao Hao, Nevena Lazic, Yasin Abbasi-Yadkori, Pooria Joulani, and Csaba Szepesvari. Provably efficient adaptive approximate policy iteration. arXiv preprint arXiv:2002.03069, 2020.

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.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. Journal of Machine Learning Research, 11(Apr):1563–1600, 2010.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is Q-learning provably efficient? In Advances in Neural Information Processing Systems, pages 4863–4873, 2018.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In Conference on Learning Theory, pages 2137–2143. PMLR, 2020.

Sammie Katt, Frans Oliehoek, and Christopher Amato. Bayesian reinforcement learning in factored pomdps. arXiv preprint arXiv:1811.05612, 2018.

Michael Jong Kim. Thompson sampling for stochastic control: The finite parameter case. IEEE Transactions on Automatic Control, 62(12):6415–6422, 2017.

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

Sahin Lale, Kamyar Azizzadenesheli, Babak Hassibi, and Anima Anandkumar. Explore more and improve regret in linear quadratic regulators. arXiv preprint arXiv:2007.12291, 2020a.

Sahin Lale, Kamyar Azizzadenesheli, Babak Hassibi, and Anima Anandkumar. Logarithmic regret bound in partially observable linear dynamical systems. arXiv preprint arXiv:2003.11227, 2020b.

Miao Liu, Xuejun Liao, and Lawrence Carin. The infinite regionalized policy representation. In ICML, 2011.

Miao Liu, Xuejun Liao, and Lawrence Carin. Online expectation maximization for reinforcement learning in pomdps. In IJCAI, pages 1501–1507, 2013.

Horia Mania, Stephen Tu, and Benjamin Recht. Certainty equivalence is efficient for linear quadratic control. arXiv preprint arXiv:1902.07826, 2019.

Ian Osband and Benjamin Van Roy. Model-based reinforcement learning and the eluder dimension. arXiv preprint arXiv:1406.1853, 2014.

Ian Osband, Daniel Russo, and Benjamin Van Roy. (more) efficient reinforcement learning via posterior sampling. In Advances in Neural Information Processing Systems, pages 3003–3011, 2013.

Yi Ouyang, Mukul Gagrani, and Rahul Jain. Learning-based control of unknown linear systems with thompson sampling. arXiv preprint arXiv:1709.04047, 2017a.

Yi Ouyang, Mukul Gagrani, Ashutosh Nayyar, and Rahul Jain. Learning unknown markov decision processes: A thompson sampling approach. In Advances in Neural Information Processing Systems, pages 1333–1342, 2017b.

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. In NIPS, pages 1225–1232, 2007.

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

Guy Shani, Ronen I Brafman, and Solomon E Shimony. Model-based online learning of pomdps. In European Conference on Machine Learning, pages 353–364. Springer, 2005.

Max Simchowitz and Dylan Foster. Naive exploration is optimal for online lqr. In International Conference on Machine Learning, pages 8937–8948. PMLR, 2020.

Malcolm Strens. A bayesian framework for reinforcement learning. In ICML, volume 2000, pages 943–950, 2000.

Jayakumar Subramanian, Amit Sinha, Raihan Seraj, and Aditya Mahajan. Approximate information state for approximate planning and reinforcement learning in partially observed systems. arXiv preprint arXiv:2010.08843, 2020.

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.

Anastasios Tsiamis and George Pappas. Online learning of the kalman filter with logarithmic regret. arXiv preprint arXiv:2002.05141, 2020.

Ruosong Wang, Russ R Salakhutdinov, and Lin Yang. Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension. Advances in Neural Information Processing Systems, 33, 2020.

Tianhao Wang, Dongruo Zhou, and Quanquan Gu. Provably efficient reinforcement learning with linear function approximation under adaptivity constraints. arXiv preprint arXiv:2101.02195, 2021.

Chen-Yu Wei, Mehdi Jafarnia-Jahromi, Haipeng Luo, Hiteshi Sharma, and Rahul Jain. Model-free reinforcement learning in infinite-horizon average-reward markov decision processes. In International Conference on Machine Learning, pages 10170–10180. PMLR, 2020.

Chen-Yu Wei, Mehdi Jafarnia-Jahromi, Haipeng Luo, and Rahul Jain. Learning infinite-horizon average-reward mdps with linear function approximation. International Conference on Artificial Intelligence and Statistics, 2021.

Yi Xiong, Ningyuan Chen, Xuefeng Gao, and Xiang Zhou. Sublinear regret for learning pomdps. arXiv preprint arXiv:2107.03635, 2021.

Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In International Conference on Machine Learning, 2019.

Zihan Zhang and Xiangyang Ji. Regret minimization for reinforcement learning by evaluating the optimal bias function. In Advances in Neural Information Processing Systems, 2019.

A Regret Decomposition

Lemma A.1. R_T can be decomposed as $R_T = HE[K_T] + R_1 + R_2 + R_3$, where

$$R_1 := E\left[\sum_{k=1}^{K_T} \left(T_k - J(\pi_k)\right) - J(\pi^*)\right];$$

$$R_2 := HE\left[\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\left|\sum_{s^0} P(s^0|s_t,a_t)\left(\nu_k(s^0|s_t,a_t) - \sum_s P(s|h_t(s,\cdot),a_t)\,\nu_k(h_t(s,\cdot),a_t)\right)\right|\right];$$

$$R_3 := E\left[\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\left(c(h_t(\cdot,\cdot),a_t) - c(h_t(\cdot,\pi_k),a_t)\right)\right].$$

Proof. First, note that $E[C(s_t,a_t)|F_{t+}] = c(h_t(\cdot,\cdot),a_t)$ for any $t \geq 1$. Thus, we can write:

$$R_T = E\left[\sum_{t=1}^{T} C(s_t,a_t) - J(\pi^*)\right] = E\left[\sum_{t=1}^{T} c(h_t(\cdot,\cdot),a_t) - J(\pi^*)\right].$$

During episode k, by the Bellman equation for the sampled POMDP ν_k and that $a_t = \pi_k(h_t(\cdot,\pi_k),\pi_k)$, we can write:

$$c(h_t(\cdot,\pi_k),a_t) - J(\pi_k) = v(h_t(\cdot,\pi_k),\pi_k) - \sum_{o\in\mathcal{O}} P(o|h_t(\cdot,\pi_k),a_t,\nu_k)\,v(h^0(\cdot,\pi_k),\pi_k),$$

where $h^0 = (h_t(\cdot,\pi_k),a_t,o,\pi_k)$. Using this equation, we proceed by decomposing the regret as

$$\begin{aligned}
R_T &= E\left[\sum_{t=1}^{T} c(h_t(\cdot,\cdot),a_t) - J(\pi^*)\right] \\
&= E\left[\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1} c(h_t(\cdot,\cdot),a_t) - J(\pi^*)\right] \\
&= E\left[\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1} v(h_t(\cdot,\pi_k),\pi_k) - v(h_{t+1}(\cdot,\pi_k),\pi_k)\right] + E\left[\sum_{k=1}^{K_T} T_k - J(\pi^*)\right] \\
&\quad + E\left[\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1} v(h_{t+1}(\cdot,\pi_k),\pi_k) - \sum_{o\in\mathcal{O}} P(o|h_t(\cdot,\pi_k),a_t,\nu_k)\,v(h^0(\cdot,\pi_k),\pi_k)\right] \\
&\quad + E\left[\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1} c(h_t(\cdot,\cdot),a_t) - c(h_t(\cdot,\pi_k),a_t)\right]
\end{aligned}$$

(telescopic sum; $= :R_1$; $= :R_2'$; $= :R_3$)

where K_T is the number of episodes upto time T, t_k is the start time of episode k (we let $t_k = T+1$ for all $k > K_T$). The telescopic sum is equal to $v(h_{t_k}(\cdot,\pi_k),\pi_k) - v(h_{t_{k+1}}(\cdot,\pi_k),\pi_k) \leq H$. Thus, the first term on the right hand side is upper bounded by $HE[K_T]$. Suffices to show that $R_2' \leq R_2$. Throughout the proof, we change the order of expectation and summation at several points. A rigorous proof for why this is allowed in the case that K_T and t_k are random variables is presented in the proof of Lemma B.3.

We proceed by bounding the term R_2'. Recall that $h^0 = (h_t(\cdot,\pi_k),a_t,o,\pi_k)$ and $h_{t+1}(\cdot,\pi_k) = (h_t(\cdot,\pi_k),a_t,o_{t+1},\pi_k)$. Conditioned on $F_{t,\cdot,\pi_k}$, the only random variable in $h_{t+1}(\cdot,\pi_k)$ is o_{t+1} ($a_t = \pi_k(h_t(\cdot,\pi_k),\pi_k)$ is measurable with respect to the sigma algebra generated by F_{t,π_k}). Therefore,

$$E\left[v(h_{t+1}(\cdot,\pi_k),\pi_k)|F_{t,\pi_k}\right] = \sum_{o\in\mathcal{O}} v(h^0,\pi_k)P(o_{t+1}=o|F_{t,\pi_k}): \tag{16}$$

We claim that $P(o_{t+1} = o|F_t, \kappa) = P(o|h_t(;), a_t;)$: by the total law of probability and that $P(o_{t+1} = o|s_{t+1} = s', F_t, \kappa) = (o|s')$, we can write

$$P(o_{t+1} = o|F_t, \kappa) = \sum_{s'} (o|s') P(s_{t+1} = s'|F_t, \kappa):$$

Note that

$$\begin{aligned} P(s_{t+1} = s'|F_t, \kappa) &= \sum_{s} P(s_{t+1} = s'|s_t = s, F_t, a_t, \kappa) P(s_t = s|F_t, \kappa) \\ &= \sum_{s} (s'|s, a_t) P(s_t = s|F_t): \end{aligned}$$

Thus,

$$P(o_{t+1} = o|F_t, \kappa) = \sum_{s,s'} (o|s')(s'|s, a_t) h_t(s;) = P(o|h_t(;), a_t;): \tag{17}$$

Combining (17) with (16) and substituting into R_2^c, we get

$$R_2^0 = E\left[\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \sum_{o \in \mathcal{O}} P(o|h_t(;), a_t;) \quad P(o|h_t(;\kappa), a_t;\kappa) \quad v(h^0;\kappa) \quad :\right]$$

Recall that for any $\in$, $P(o|h_t(;), a_t;) = \sum_{s'} (o|s') \sum_{s} (s'|s, a_t) h_t(s;)$. Thus,

$$\begin{aligned} R_2^0 = E\Bigg[&\sum_{t=t_k}^{t_{k+1}-1} \sum_{o,s'} v(h^0;\kappa)(o|s') \sum_{s} (s'|s, a_t) h_t(s;) \Bigg] \\ &E\Bigg[\sum_{t=t_k}^{t_{k+1}-1} \sum_{o,s'} v(h^0;\kappa)(o|s') \sum_{s} {}_\kappa(s'|s, a_t) h_t(s;) \Bigg] \\ &+ E\Bigg[\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \sum_{o,s'} v(h^0;\kappa)(o|s') \sum_{s} {}_\kappa(s'|s, a_t) \quad h_t(s;) \quad h_t(s;\kappa) \quad :\Bigg] \end{aligned} \tag{18}$$

For the first term, note that conditioned on $F_t;$, the distribution of s_t is $h_t(;)$ by the definition of h_t. Furthermore, a_t is measurable with respect to the sigma algebra generated by $F_t;\kappa$ since $a_t = (h_t(;\kappa);\kappa)$. Thus, we have

$$E\left[v(h^0;\kappa) \sum_{s} (s'|s, a_t) h_t(s;) F_t;\kappa\right] = v(h^0;\kappa) E\left[(s'|s_t, a_t) F_t;\kappa\right]: \tag{19}$$

Similarly, for the second term on the right hand side of (18), we have

$$E\left[v(h^0;\kappa) \sum_{s} {}_\kappa(s'|s, a_t) h_t(s;) F_t;\kappa\right] = v(h^0;\kappa) E\left[{}_\kappa(s'|s_t, a_t) F_t;\kappa\right]: \tag{20}$$

Replacing (19), (20) into (18) and using the tower property of conditional expectation, we get

$$\begin{aligned} R_2^0 = E\Bigg[&\sum_{t=t_k}^{t_{k+1}-1} \sum_{s'} \sum_{o} v(h^0;\kappa)(o|s') \quad (s'|s_t, a_t) \quad {}_\kappa(s'|s_t, a_t) \Bigg] \\ &+ E\Bigg[\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \sum_{s'} \sum_{o} v(h^0;\kappa)(o|s') \sum_{s} {}_\kappa(s'|s, a_t) \quad h_t(s;) \quad h_t(s;\kappa) \quad :\Bigg] \end{aligned} \tag{21}$$

Since $\sup_{b \in} \sum_{s} v(b;\kappa) \quad H$ and $\sum_{o} (o|s') = 1$, the inner summation for the first term on the right hand side of (21) can be bounded as

$$\sum_{o \in \mathcal{O}} v(h^0;\kappa)(o|s') \quad (s'|s_t, a_t) \quad {}_\kappa(s'|s_t, a_t) \quad H(s'|s_t, a_t) \quad {}_\kappa(s'|s_t, a_t): \tag{22}$$

Using $\sup_{b2\ s} v(b;k)\ H$, $\sum_o (ojs^0) = 1$ and $\sum_s {}_{0\,k}(s^0js;a_t) = 1$, the second term on the right hand side of (21) can be bounded as

$$\sum_s \sum_o v(h^0;k)(ojs^0) \sum_s {}_k(s^0js;a_t)h_t(s;) \quad h_t(s;k)\ H \sum_{s^0\ o2O} h_t(s;) \quad h_t(s;k) \tag{23}$$

Substituting (22) and (23) into (21) proves that $R_2^c\ R_2$. □

B Proofs of Section 5

B.1 Full Upper Bound on the Expected Regret of Theorem 2

The exact expression for the upper bound of the expected regret in Theorem 2 is

$$\begin{aligned}
&E[R_T] = H\,E[K_T] + E[R_1] + E[R_2] + E[R_3] \\
&(1+H)E[K_T] + 12HK_2(jSjjAjT)^{2=3} \\
&\quad + (H+1)K_1 E\left[\sum_{k=1}^{K_T} \sqrt{\frac{T_k}{t_k}}\right] + 2 + H \\
&(1+H)\sqrt{2T(1+jSjjAj\log(T+1))} + 12HK_2(jSjjAjT)^{2=3} + \\
&\quad 7(H+1)K_1\left(\sqrt{2T(1+jSjjAj\log(T+1))}\log\sqrt{2T}\right) + 2 + H:
\end{aligned}$$

B.2 Finite-parameter Case Satisfies Assumptions 3 and 4

In this section, we show that Assumptions 3 and 4 are satisfied for the finite-parameter case i.e., $jj < 1$ as long as the PSRL-POMDP generates a deterministic schedule. As an instance, a deterministic schedule can be generated by choosing $\hat{m}_t(s;a) = t$ for all state-action pairs $(s;a)$ and running Algorithm 1 with either $\text{SCHED}(t_k;T_{k\ 1}) = 2t_k$ or $\text{SCHED}(t_k;T_{k\ 1}) = t_k + T_{k\ 1}$.

Lemma B.1. Assume $jj < 1$. If Algorithm 1 generates a deterministic schedule, then Assumptions 3 and 4 are satisfied.

Proof. Observe that the left hand side of (10) is zero if $_{k(t)} = $, and is upper bounded by 2 if $_{k(t)} = $. Thus, we can write

$$E\left[\sum_s h_t(s;) \quad h_t(s;_{k(t)}) \quad 2\right] P(_{k(t)} = j) = 2E\left[1 \quad f_{t\,_{k(t)}}()j \quad \exp(\ t_{k(t)})\right];$$

which obviously satisfies Assumption 3 by choosing a large enough constant K_1. Here, the last equality is by Lemma 2 and that the start time of episode k(t) is deterministic.

To see why Assumption 4 is satisfied, let $\hat{}_t$ be the Maximum a Posteriori (MAP) estimator, i.e., $\hat{}_t = \operatorname{argmax}_{2} f_t()$. Then, the left hand side of (11) is equal to zero if $\hat{}_t = $. Note that this happens with high probability with the following argument:

$$P(\hat{}_t = j)\ \ P(f_t()\ \ 0{:}5j) = P(1 \quad f_t()\ \ 0{:}5j)\ \ 2E[1 \quad f_t()j]\ \ 2\exp(\ t):$$

Here the first inequality is by the fact that if $f_t() > 0{:}5$, then the MAP estimator would choose $_t = \hat{}$. The second inequality is by applying Markov inequality and the last inequality is by Lemma 2. Note that $\hat{m}_t(s;a)\ \ t$ by definition. We claim that Assumption 4 is satisfied by choosing $K_2 = 2\ (\ 1=)\log(=2)$. To see this, note that $2\exp(\ t)\ \ $ for $t\ (\ 1=)\log(=2)$. In this case, (11) automatically holds since with probability at least $1\ \ $ the left hand side is zero. For $t < (\ 1=)\log(=2)$, note that the left hand side of (11) can be at most 2. Therefore, K_2 can be found by solving $2\ K_2 = \ (\ 1=)\log(=2)$. □

B.3 Auxiliary Lemmas for Section 5

Lemma B.2. [Lemma 3 in Ouyang et al. [2017b]] The term $E[R_1]$ can be bounded as $E[R_1] \leq E[K_T]$.

Proof.

$$E[R_1] = E\Big[\sum_{k=1}^{K_T} \big[T_k J(k) - J()\big]\Big] = E\Big[\sum_{k=1}^{\infty} 1(t_k \leq T) T_k J(k)\Big] - TE[J()].$$

By monotone convergence theorem and the fact that $J(k) \leq 0$ and $T_k \leq T_{k-1}+1$ (the first criterion in determining the episode length in Algorithm 1), the first term can be bounded as

$$E\Big[\sum_{k=1}^{\infty} 1(t_k \leq T) T_k J(k)\Big] = \sum_{k=1}^{\infty} E\big[1(t_k \leq T) T_k J(k)\big] \leq \sum_{k=1}^{\infty} E\big[1(t_k \leq T)(T_{k-1}+1) J(k)\big].$$

Note that $1(t_k \leq T)(T_{k-1}+1)$ is $\mathcal{F}_{t_k}$-measurable. Thus, by the property of posterior sampling (Lemma 3), $E[1(t_k \leq T)(T_{k-1}+1)J(k)] = E[1(t_k \leq T)(T_{k-1}+1)J()]$. Therefore,

$$\begin{aligned} E[R_1] &\leq E\Big[\sum_{k=1}^{\infty} 1(t_k \leq T)(T_{k-1}+1) J()\Big] - TE[J()] \\ &= E\big[J()\big(K_T + \sum_{k=1}^{K_T} T_{k-1}\big)\big] - TE[J()] \\ &= E[J()K_T] + E\Big[J()\Big(\sum_{k=1}^{K_T} T_{k-1} - T\Big)\Big] \leq E[K_T], \end{aligned}$$

where the last inequality is by the fact that $\sum_{k=1}^{K_T} T_{k-1} - T \leq 0$ and $0 \leq J() \leq 1$. □

Lemma B.3. The term $E[R_3]$ can be bounded as

$$E[R_3] \leq K_1 E\Big[\sum_{k=1}^{K_T} \sqrt{\frac{T}{t_k}}\Big] + 1,$$

where $K_1 := K_1(|S|, |A|, |O|,)$ is the constant in Assumption 3.

Proof. Recall that

$$E[R_3] = E\Big[\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \big[c(h_t(;);a_t) - c(h_t(;_k);a_t)\big]\Big].$$

Let $k(t)$ be a random variable denoting the episode number at time t, i.e., $t_{k(t)} \leq t < t_{k(t)+1}$ for all $t \leq T$. By the definition of c, we can write

$$\begin{aligned} E[R_3] &= E\Big[\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \sum_{s} C(s,a_t)\big[h_t(s;) - h_t(s;_k)\big]\Big] \\ &= E\Big[\sum_{t=1}^{T} \sum_{s} C(s,a_t)\big[h_t(s;) - h_t(s;_{k(t)})\big]\Big] \\ &\quad + \sum_{t=1}^{T} E\Big[\sum_{s} h_t(s;) - h_t(s;_{k(t)})\Big] \\ &= E\Big[\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} E\Big[\sum_{s} h_t(s;) - h_t(s;_k)\Big]\Big]. \end{aligned}$$

where the inequality is by $0 \le C(s,a_t) \le 1$. Let $K_1 := K_1(|S|,|A|,|O|,\epsilon)$ be the constant in Assumption 3 and define event E_1 as the successful event of Assumption 3 where $\mathbb{E}\left[\sum_s |\hat{h}_t(s,\cdot) - h_t(s,\cdot)|\right] \le \frac{K_1}{\sqrt{t_k}}$ happens. We can write

$$\begin{aligned}
&\mathbb{E}\left[\sum_s |\hat{h}_t(s,\cdot) - h_t(s,\cdot)|\right] \\
&= \mathbb{E}\left[\sum_s |\hat{h}_t(s,\cdot) - h_t(s,\cdot)| \left(\mathbb{1}(E_1) + \mathbb{1}(E^c)\right)\right] \\
&\le \frac{K_1}{\sqrt{t_k}} + 2\mathbb{1}(E^c) \le \frac{K_1}{\sqrt{t_k}} + \frac{2}{t_k}.
\end{aligned}$$

Recall that by Assumption 3, $P(E_1^c) \le \frac{1}{t}$. Therefore,

$$\mathbb{E}[R_3] \le K_1 \mathbb{E}\left[\sum_{k=1}^{K_T} \frac{T_k}{\sqrt{t_k}}\right] + 2T.$$

Choosing $\epsilon = \min(1/(2T), 1/(2HT))$ completes the proof. □

Lemma B.4. The term R_2 can be bounded as

$$R_2 \le H + 12HK_2(|S||A|T)^{2/3},$$

where $K_2 := K_2(|S|,|A|,|O|,\epsilon)$ in Assumption 4.

Proof. Recall that

$$R_2 = H\,\mathbb{E}\left[\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\sum_{s'} |\hat{\mathbb{P}}_{t_k}(s'|s_t,a_t) - \mathbb{P}_k(s'|s_t,a_t)|\right]. \tag{24}$$

We proceed by bounding the inner term of the above equation. For notational simplicity, define $z := (s,a)$ and $z_t := (s_t,a_t)$. Let $\hat{\mathbb{P}}_{t_k}$ be the estimator in Assumption 4 and define the confidence set $\mathcal{B}_k$ as

$$\mathcal{B}_k := \left\{\mathbb{P} : \sum_{s'\in S} |\mathbb{P}(s'|z) - \hat{\mathbb{P}}_{t_k}(s'|z)| \le \frac{K_2\sqrt{8z\ln S}}{\sqrt{\max\{1, \hat{m}_{t_k}(z)\}}}, \forall z \in S\times A\right\};$$

where $K_2 := K_2(|S|,|A|,|O|,\epsilon)$ is the constant in Assumption 4. Note that $\mathcal{B}_k$ reduces to the confidence set used in Jaksch et al. [2010], Ouyang et al. [2017b] in the case of perfect observation by choosing $\hat{m}_t(s,a) = n_t(s,a)$. By triangle inequality, the inner term in (24) can be bounded by

$$\begin{aligned}
&\sum_{s'} |\mathbb{P}(s'|z_t) - \mathbb{P}_k(s'|z_t)| \\
&\le \sum_{s'} |\mathbb{P}(s'|z_t) - \hat{\mathbb{P}}_{t_k}(s'|z_t)| + \sum_{s'} |\mathbb{P}_k(s'|z_t) - \hat{\mathbb{P}}_{t_k}(s'|z_t)| \\
&\le 2\,\mathbb{1}(\mathbb{P} \notin \mathcal{B}_k) + \mathbb{1}(\mathbb{P}_k \notin \mathcal{B}_k) + \frac{2K_2}{\sqrt{\max\{1, \hat{m}_{t_k}(z_t)\}}}.
\end{aligned}$$

Substituting this into (24) implies

$$R_2 \le 2H\,\mathbb{E}\left[\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1} \mathbb{1}(\mathbb{P} \notin \mathcal{B}_k) + \mathbb{1}(\mathbb{P}_k \notin \mathcal{B}_k)\right] + 2H\,\mathbb{E}\left[\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1} \frac{K_2}{\sqrt{\max\{1, \hat{m}_{t_k}(z_t)\}}}\right]. \tag{25}$$

We need to bound these two terms separately.

Bounding the first term. For the first term we can write:

$$E\left[\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\frac{1}{\sqrt{N_{t_k}(z_t)}}\left(\mathbb{1}(\hat{\mu}_k \geq B_k) + \mathbb{1}(\mu_k \geq B_k)\right)\right] = E\left[\sum_{k=1}^{T} T_k\left(\mathbb{1}(\hat{\mu}_k \geq B_k) + \mathbb{1}(\mu_k \geq B_k)\right)\right]$$
$$T\,E\left[\sum_{k=1}^{K_T}\left(\mathbb{1}(\hat{\mu}_k \geq B_k) + \mathbb{1}(\mu_k \geq B_k)\right)\right]$$
$$T\sum_{k=1}^{T} E\left[\left(\mathbb{1}(\hat{\mu}_k \geq B_k) + \mathbb{1}(\mu_k \geq B_k)\right)\right];$$

where the last inequality is by the fact that $K_T \leq T$. Now, observe that since B_k is $\mathcal{F}_{t_k}$-measurable, Lemma 3 implies that $E[\mathbb{1}(\mu_k \geq B_k)] = E[\mathbb{1}(\hat{\mu}_k \geq B_k)]$. Moreover, by Assumption 4, $E[\mathbb{1}(\hat{\mu}_k \geq B_k)] = P(\hat{\mu}_k \geq B_k)$. By choosing $\epsilon = \frac{1}{4T^2}$, we get

$$E\left[\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\frac{1}{\sqrt{N_{t_k}(z_t)}}\left(\mathbb{1}(\hat{\mu}_k \geq B_k) + \mathbb{1}(\mu_k \geq B_k)\right)\right] \leq \frac{1}{2}. \tag{26}$$

Bounding the second term. To bound the second term of (25), observe that by the second criterion of the algorithm in choosing the episode length, we have $2\hat{m}_{t_k}(z_t) \leq \hat{m}_t(z_t)$. Thus,

$$E\left[\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\frac{K_2}{\sqrt{\max\{1, \hat{m}_{t_k}(z_t)\}}}\right] \leq E\left[\sum_{t=1}^{T}\frac{\sqrt{2}K_2}{\sqrt{\max\{1, \hat{m}_t(z_t)\}}}\right]$$
$$= \sum_{t=1}^{T}\sum_{z} E\left[\frac{\sqrt{2}K_2\mathbb{1}(z_t = z)}{\sqrt{\max\{1, \hat{m}_t(z)\}}}\right]$$
$$= \sum_{t=1}^{T}\sum_{z} E\left[\frac{\sqrt{2}K_2\mathbb{1}(z_t = z)}{\sqrt{\max\{1, \hat{m}_t(z)\}}}\mathbb{1}\,\hat{m}_t(z) \geq n_t(z)\right]$$
$$+ \sum_{t=1}^{T}\sum_{z} E\left[\frac{\sqrt{2}K_2\mathbb{1}(z_t = z)}{\sqrt{\max\{1, \hat{m}_t(z)\}}}\mathbb{1}\,\hat{m}_t(z) < n_t(z)\right]$$
$$\leq \sum_{t=1}^{T}\sum_{z} E\left[\frac{\sqrt{2}K_2\mathbb{1}(z_t = z)}{\sqrt{\max\{1, n_t(z)\}}}\right] + \sum_{t=1}^{T}\sum_{z} E\left[\sqrt{2}K_2\mathbb{1}\,\hat{m}_t(z) < n_t(z)\right]. \tag{27}$$

Lemma 4 implies that $E[\mathbb{1}\,\hat{m}_t(z) < n_t(z)] = P(\hat{m}_t(z) < n_t(z))$. Thus, the second term in (27) can be bounded by $\sqrt{2}K_2|S||A|T$. To bound the first term of (27), we can write:

$$\sum_{t=1}^{T}\sum_{z} E\left[\frac{\sqrt{2}K_2\mathbb{1}(z_t = z)}{\sqrt{\max\{1, n_t(z)\}}}\right]$$
$$\sqrt{2}K_2\,E\left[\sum_{z}\sum_{t=1}^{N_T(z)}\frac{\mathbb{1}(z_t = z)}{\sqrt{\max\{1, n_t(z)\}}}\right]:$$

Observe that whenever $z_t = z$, $n_t(z)$ increases by 1. Since, $n_t(z)$ is the number of visits to z by time $t-1$ (including $t-1$ and excluding t), the denominator will be 1 for the first two times that $z_t = z$. Therefore, the term inside the expectation can be bounded by

$$\sum_{z}\sum_{t=1}^{T}\frac{\mathbb{1}(z_t = z)}{\sqrt{\max\{1, n_t(z)\}}} = \sum_{z} E\left[\mathbb{1}(n_{T+1}(z) > 0) + \sum_{j=1}^{n_{T+1}(z)-1}\frac{1}{\sqrt{j}}\right]$$
$$\sum_{z} E\left[\mathbb{1}(n_{T+1}(z) > 0) + 2\sqrt{n_{T+1}(z)}\right]$$
$$3\sum_{z}\sqrt{n_{T+1}(z)}:$$

Since $\sum_z n_{T+1}(z) = T$, Cauchy Schwartz inequality implies

$$3 \sum_z \sqrt{n_{T+1}(z)} \le 3 \sqrt{|S||A| \sum_z n_{T+1}(z)} = 3\sqrt{|S||A|T}.$$

Therefore, the first term of (27) can be bounded by

$$\sum_{t=1}^{T} \sum_z \mathbb{E}\left[\frac{\sqrt{2}K_2 \mathbb{1}(z_t = z)}{\sqrt{\max\{1, n_t(z)\}}}\right] \le 3K_2 \sqrt{2|S||A|T}.$$

Substituting this bound in (27) along with the bound on the second term of (27), we obtain

$$\mathbb{E}\left[\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \frac{K_2}{\sqrt{\max\{1, \hat{m}_{t_k}(z_t)\}}}\right] \le 3K_2 \sqrt{2|S||A|T} + \sqrt{2}K_2|S||A|T.$$

$= (3/2)^{2/3}(|S||A|T)^{1/3}$ minimizes the upper bound, and thus

$$\mathbb{E}\left[\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \frac{K_2}{\sqrt{\max\{1, \hat{m}_{t_k}(z_t)\}}}\right] \le 6K_2(|S||A|T)^{2/3}. \tag{28}$$

By substituting (26) and (28) into (25), we get

$$R_2 \le H + 12HK_2(|S||A|T)^{2/3}.$$

□

Lemma B.5. The following inequalities hold:

1. The number of episodes K_T can be bounded as $K_T \le \sqrt{2T(1 + |S||A|)\log(T + 1)} = O(\sqrt{|S||A|T})$.
2. The following inequality holds: $\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \frac{1}{\sqrt{\max\{1, \hat{m}_{t_k}(z_t)\}}} \le \sqrt{2T(1 + |S||A|)\log(T + 1)}\log\sqrt{2T} = \tilde{O}(|S||A|\sqrt{T})$.

Proof. We first provide an intuition why these results should be true. Note that the length of the episodes is determined by two criteria. The first criterion triggers when $T_k = T_{k-1} + 1$ and the second criterion triggers when the pseudo counts doubles for a state-action pair compared to the beginning of the episode. Intuitively speaking, the second criterion should only happen logarithmically, while the first criterion occurs more frequently. This means that one could just consider the first criterion for an intuitive argument. Thus, if we ignore the second criterion, we get $T_k = O(k)$, $K_T = O(\sqrt{T})$, and $t_k = O(k^2)$ which implies $\sum_{k=1}^{K_T} \frac{T_k}{\sqrt{t_k}} = O(K_T) = O(\sqrt{T})$. The rigorous proof is stated in the following.

1. Define macro episodes with start times t_{m_i} given by $t_{m_1} = t_1$ and $t_{m_{i+1}} :=$

$$\min\{t_k > t_{m_i} : \hat{m}_{t_k}(s, a) > 2\hat{m}_{t_{k-1}}(s, a) \text{ for some } (s, a)\}.$$

Note that a new macro episode starts when the second criterion of episode length in Algorithm 1 triggers. Let M_T be the random variable denoting the number of macro episodes by time T and define $m_{M_T+1} = K_T + 1$.

Let $\tilde{T}_i$ denote the length of macro episode i. Note that $\tilde{T}_i = \sum_{k=m_i}^{m_{i+1}-1} T_k$. Moreover, from the definition of macro episodes, we know that all the episodes in a macro episode except the last one are triggered by the first criterion, i.e., $T_k = T_{k-1} + 1$ for all $m_i \le k \le m_{i+1} - 2$. This implies that

$$\tilde{T}_i = \sum_{k=m_i}^{m_{i+1}-1} T_k = T_{m_{i+1}-1} + \sum_{j=1}^{m_{i+1}-m_i-1} (T_{m_i-1} + j)$$

$$\le 1 + \sum_{j=1}^{m_{i+1}-m_i-1} (1 + j) = \frac{(m_{i+1} - m_i)(m_{i+1} - m_i + 1)}{2}.$$

This implies that $m_{i+1} \leq m_i + \sqrt{2T_i}$. Now, we can write:

$$K_T = m_{M_T+1} - 1 = \sum_{i=1}^{M_T} (m_{i+1} - m_i)$$
$$\leq \sum_{i=1}^{M_T} \sqrt{2T_i} = \sqrt{2M_T} \sum_i \sqrt{T_i} = \sqrt{2M_T T}; \tag{29}$$

where the last inequality is by Cauchy-Schwartz.

Now, it suffices to show that $M_T \leq 1 + |S||A| \log(T+1)$. Let $T_{s,a}$ be the start times at which the second criterion is triggered at state-action pair (s,a), i.e.,

$$T_{s,a} := \{t_k \leq T : m_{t_k}(s,a) > 2m_{t_{k-1}}(s,a)\}.$$

We claim that $|T_{s,a}| \leq \log(m_{T+1}(s,a))$. To prove this claim, assume by contradiction that $|T_{s,a}| \geq \log(m_{T+1}(s,a)) + 1$, then

$$m_{t_{K_T}}(s,a) \prod_{t_k \leq T; m_{t_k}(s,a) \geq 1} \frac{m_{t_k}(s,a)}{m_{t_k-1}(s,a)} \prod_{t_k \in T_{s,a}; m_{t_k-1}(s,a) \geq 1} \frac{m_{t_k}(s,a)}{m_{t_k-1}(s,a)}$$
$$> \prod_{t_k \in T_{s,a}; m_{t_k-1}(s,a) \geq 1} 2 = 2^{|T_{s,a}|-1} \geq m_{T+1}(s,a);$$

which is a contradiction. The second inequality is by the fact that $m_t(s,a)$ is non-decreasing, and the third inequality is by the definition of $T_{s,a}$. Therefore,

$$M_T \leq 1 + \sum_{s,a} |T_{s,a}| \leq 1 + \sum_{s,a} \log(m_{T+1}(s,a))$$
$$\leq 1 + |S||A| \log\Big(\sum_{s,a} m_{T+1}(s,a)/|S||A|\Big)$$
$$= 1 + |S||A| \log(T+1); \tag{30}$$

where the third inequality is due to the concavity of log and the last inequality is by the fact that $m_{T+1}(s,a) \leq T+1$.

2. First, we claim that $T_k \leq \sqrt{2T}$ for all $k \leq K_T$. To see this, assume by contradiction that $T_k > \sqrt{2T}$ for some $k \leq K_T$. By the first stopping criterion, we can conclude that $T_{k-1} > \sqrt{2T} - 1$, $T_{k-2} > \sqrt{2T} - 2$, ..., $T_1 > \max\{\sqrt{2T} - k + 1; 0\}$ since the episode length can increase at most by one compared to the previous one. Note that $k \leq \sqrt{2T} - 1$, because otherwise $T_1 > 2$ which is not feasible since $T_1 \leq T_0 + 1 = 2$. Thus, $\sum_{k=1}^{k} T_k > 0.5\sqrt{2T}(\sqrt{2T}+1) > T$ which is a contradiction.

We now proceed to lower bound t_k. By the definition of macro episodes in part (1), during a macro episode length of the episodes except the last one are determined by the first criterion, i.e., for macro episode i, one can write $T_k = T_{k-1} + 1$ for $m_i \leq k \leq m_{i+1} - 2$. Hence, for $m_i \leq k \leq m_{i+1} - 2$,

$$t_{k+1} = t_k + T_k = t_k + T_{m_i-1} + k - (m_i - 1)$$
$$\geq t_k + k - m_i + 1.$$

Recursive substitution of t_k implies that $t_k \geq t_{m_i} + 0.5(k - m_i)(k - m_i + 1)$ for $m_i \leq k \leq m_{i+1} - 1$. Thus,

$$\sum_{k=1}^{K_T} \frac{T_k}{\sqrt{t_k}} \leq \sqrt{2T} \sum_{i=1}^{M_T} \sum_{k=m_i}^{m_{i+1}-1} \frac{1}{\sqrt{t_k}}$$
$$\leq \sqrt{2T} \sum_{i=1}^{M_T} \sum_{k=m_i}^{m_{i+1}-1} \frac{1}{\sqrt{t_{m_i} + 0.5(k - m_i)(k - m_i + 1)}}. \tag{31}$$

The denominator of the summands at $k = m_i$ is equal to $\sqrt{t_{m_i}}$. For other values of k it can be lower bounded by $0.5(k - m_i)^2$. Thus,

$$\begin{aligned}
&\sum_{i=1}^{M_T} \sum_{k=m_i}^{m_{i+1}-1} \frac{1}{\sqrt{t_{m_i} + 0.5(k - m_i)(k - m_i + 1)}} \\
&\le \sum_{i=1}^{M_T} \frac{1}{\sqrt{t_{m_i}}} + \sum_{i=1}^{M_T} \sum_{k=m_i+1}^{m_{i+1}-1} \frac{2}{\sqrt{k - m_i}} \\
&\le M_T + \sum_{i=1}^{M_T} \sum_{j=1}^{m_{i+1}-m_i-1} \frac{2}{\sqrt{j}} \\
&\le M_T + 2(M_T + \sum_{i=1}^{M_T} \log(m_{i+1} - m_i)) \\
&\le M_T(1 + 2) + 2M_T \log\left(\frac{1}{M_T} \sum_{i=1}^{M_T} (m_{i+1} - m_i)\right) \\
&\le M_T(1 + 2) + 2M_T \log \frac{2T}{M_T} \\
&\le 7M_T \log \frac{2T}{M_T},
\end{aligned}$$

where the second inequality is by $t_{m_i} \ge 1$, the third inequality is by the fact that $\sum_{j=1}^{K-1} 1/j \le 1 + \int_1^K dx/x = 1 + \log K$, the forth inequality is by concavity of log and the fifth inequality is by the fact that $\sum_{i=1}^{M_T} (m_{i+1} - m_i) = m_{M_T+1} - 1 = K_T$ and $K_T/M_T \le 2T/M_T \le 2T$ (see (29)). Substituting this bound into (31) and using the upper bound on M_T (30), we can write

$$\sum_{k=1}^{K_T} \frac{1}{\sqrt{t_k}} \le 7M_T \log \frac{2T}{M_T} \le 7\sqrt{2T}(1 + |S||A| \log(T + 1)) \log \sqrt{2T}.$$

□

C Other Proofs

C.1 Proof of Lemma 1

Lemma (restatement of Lemma 1). *Suppose Assumption 1 holds. Then, the policy $(\cdot;\cdot) : \mathcal{S} \to \mathcal{A}$ given by*

$$(b;\theta) := \operatorname{argmin}\{c(b,a) + \sum_{o \in \mathcal{O}} P(o|b,a;\theta) v(b^o;\theta)\}_{a \in \mathcal{A}}$$

is the optimal policy with $J(h;\theta) = J(\theta)$ for all $h \in \mathcal{S}$.

Proof. We prove that for any policy π, $J(h;\pi) - J(h;\theta) = J(\pi) - J(\theta)$ for all $h \in \mathcal{S}$. Let $\pi : \mathcal{S} \to \mathcal{A}$ be an

arbitrary policy. We can write

$$
\begin{aligned}
J(h;\pi) &= \limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} E[C(s_t,\pi(h_t))|s_1 = h] \\
&= \limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} E\Big[E[C(s_t,\pi(h_t))|F_t, s_1 = h]\,\Big|\, s_1 = h\Big] \\
&= \limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} E[c(h_t,\pi(h_t))|s_1 = h] \\
&\quad \limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} E[J(\pi) + v(h_t,\pi) - v(h_{t+1},\pi)|s_1 = h] \\
&= J(\pi);
\end{aligned}
$$

with equality attained by π^* completing the proof. □