Posterior Sampling for Continuing Environments

Wanqiao Xu
Stanford University

Shi Dong
Google DeepMind

Benjamin Van Roy
Stanford University

Abstract

Existing posterior sampling algorithms for continuing reinforcement learning (RL) rely on maintaining state-action visitation counts, making them unsuitable for complex environments with high-dimensional state spaces. We develop the first extension of posterior sampling for RL (PSRL) that is suited for a continuing agent-environment interface and integrates naturally into scalable agent designs. Our approach, Continuing PSRL, determines when to resample a new model of the environment from the posterior distribution based on a simple randomization scheme. We establish an $\tilde{O}(\tau S\sqrt{AT})$ bound on the Bayesian regret in the tabular setting, where S is the number of environment states, A is the number of actions, and τ denotes the *reward averaging time*, which is a bound on the duration required to accurately estimate the average reward of any policy. Our work is the first to formalize and rigorously analyze this random resampling approach. Our simulations demonstrate Continuing PSRL's effectiveness in high-dimensional state spaces where traditional algorithms fail.

1 Introduction

A reinforcement learning (RL) agent is faced with the task of interacting with an unknown environment while trying to maximize the total reward accrued over time. A core challenge in RL is how to balance the fundamental tradeoff: when taking exploratory actions, the agent accrues more knowledge about the unknown environment, but exploiting the knowledge obtained so far may result in higher immediate return.

A growing literature builds on Thompson sampling (Thompson, 1933; Russo et al., 2018) to develop randomized approaches to exploration (Osband et al., 2013; 2016b;a). While these approaches have proved to be effective, they have largely been limited to episodic environments. In particular, the *modus oparandi* involves randomly sampling a new policy, which aims to maximize expected return in a statistically plausible model of the environment, immediately before the start of each episode, and following that policy throughout the episode. For example, bootstrapped DQN (Osband et al., 2016a) maintains an ensemble that approximates the posterior distribution of the optimal action value function Q_* and, before each ℓth episode, samples a random element $\hat{Q}_\ell$. Then, a greedy policy with respect to $\hat{Q}_\ell$ is executed.

While learning in continuing environments is a fundamental problem in RL (Naik et al., 2019), work on randomized approaches to exploration have largely focused on episodic environments, with few exceptions specific to a *tabula rasa* context (Ouyang et al., 2017; Theocharous et al., 2018). We develop for the first time a version of posterior sampling for reinforcement learning (PSRL) that

is easily extended to work with function approximation in complex environments. Our proposed method, Continuing PSRL, resamples a new policy at each time with a probability p. Here, p is specified as a part of the agent design and represents how the agent chooses to partition its experience into intervals that it interprets as trials. With this resampling rule, it is natural to execute a policy that maximizes discounted return with a discount factor $\gamma = 1 - p$. Indeed, a simple lemma shows that, with this choice of γ, the undiscounted return attained between consecutive resampling times constitutes an unbiased estimate of the γ-discounted return of the policy used. This simple resampling scheme easily integrates into randomized exploration algorithms with function approximation. For example, bootstrapped DQN (Osband et al., 2016a) can be modified to address continuing environments by resampling the action value function from the prevailing approximate posterior distribution at each time with probability p. As with the original version of bootstrapped DQN, each executed action is greedy with respect to the most recently sampled action value function.

Many theoretical works consider γ as part of the environment, e.g. they directly analyze γ-discounted regret (Lattimore & Hutter, 2012; Wang et al., 2020). In contrast, we assess agent performance in terms of undiscounted regret. Thus, while the discount factor γ plays a role in agent design, it does not reflect what we view as the designer's objective. This viewpoint aligns with empirical work that regards the discount factor as a tunable hyperparameter (Mnih et al., 2015; François-Lavet et al., 2015). Our analysis shows that, while resampling with a probability $p = 1 - \gamma$ and planning with a γ-discounted objective does not lead to vanishing per-timestep regret, that can be accomplished by increasing γ over time.

Prior works considered versions of PSRL that treat continuing environments (Ouyang et al., 2017; Theocharous et al., 2018) by directly planning under the undiscounted regret. The algorithm proposed in Ouyang et al. (2017) resamples an environment from the environment posterior each time either of the two following criteria holds: 1) the time elapsed since the last resampling exceeds the interval between the two most recent resamplings, and 2) the number of visits to any state-action pair is doubled since the previous resampling. The latter criterion plays an essential role but is not viable when operating in a complex environment, for example, addressing an intractably large state space and approximating a distribution of action value functions using a neural network (Osband et al., 2016a; Dwaracherla et al., 2020). In particular, it is not clear how to efficiently track visitation counts, and even if that were possible, the counts could be irrelevant since it may even be rare to visit any individual state more than once. In order to address large state spaces, Theocharous et al. (2018) considers simply doubling the duration between each successive pair of resampling times. Although the resulting algorithm circumvents maintaining visitation counts, their analysis relies heavily on technical assumptions, without which the regret bound grows linearly with time. Another work (Tang et al., 2024) that came after the first version of our paper considered using a fixed resampling schedule that yields a prior-dependent regret bound, but their result exhibits a loose dependence on problem parameters.

In this paper, we formalize our aforementioned resampling approach to randomized exploration – both with fixed and decreasing reset probabilities – and provide a first rigorous analysis, which establishes regret bounds similar to Ouyang et al. (2017) but with a resampling criterion much simpler and more scalable than what is proposed in that paper. Interestingly, our analysis is also simpler than that of Ouyang et al. (2017) because our resampling criterion is policy-independent. Specifically, we show that for a choice of discount factor that suitably depends on the horizon T, our algorithm, continuing posterior sampling, satisfies an $\tilde{O}(\tau S\sqrt{AT})$ regret bound, where S is the number of environment states, A is the number of actions, and τ denotes the reward averaging time (Dong et al., 2022), which is a bound on the duration required to accurately estimate the average reward of any policy. This regret bound matches that established by Ouyang et al. (2017), though their use of "span" replaced by the reward averaging time τ.

2 Problem Formulation

We consider the problem of learning to optimize performance through a single stream of interactions with an unknown environment $\mathcal{E} = (\mathcal{A}, \mathcal{S}, \rho)$, modeled as a Markov decision process (MDP). Here $\mathcal{A}$ is a finite action space with cardinality A, $\mathcal{S}$ is a finite state space with cardinality S, and ρ is a function that specifies a state transition probability $\rho(s' \mid s, a)$ given a current state $s \in \mathcal{S}$ and action $a \in \mathcal{A}$. Interactions up to time t make up a history $\mathcal{H}_t = (S_0, A_0, S_1, A_1, S_2, \ldots, A_{t-1}, S_t)$, and the agent selects action A_t after observing S_t. The environment and all associated random quantities we consider are defined within a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. In particular, the environment $\mathcal{E}$ itself is random, and we use a distribution $\mathbb{P}(\mathcal{E} \in \cdot)$ to capture the agent designer's prior belief over all possible environments. As the history evolves, what can be learned is represented by the posterior distribution $\mathbb{P}(\mathcal{E} \in \cdot | \mathcal{H}_t)$. We additionally assume that $\mathcal{A}$ and $\mathcal{S}$ are deterministic and known, but the observation probability function ρ is a random variable that the agent needs to learn. For simplicity, we assume S_0 is deterministic, but the same analysis can be easily extended to consider a distribution over initial states.

An agent's preferences can be represented by a *reward function* $r : \mathcal{S} \times \mathcal{A} \mapsto [0, 1]$. After selecting action A_t in state S_t, the agent observes S_{t+1} and receives a deterministic reward $R_{t+1} = r(S_t, A_t)$ bounded in $[0, 1]$. We take r to be deterministic and known for simplicity, but our result easily generalizes to randomized reward functions. The agent specifies its actions via *policies*. A stochastic policy π can be represented by a probability mass function $\pi(\cdot \mid S_t)$ that an agent assigns to actions in $\mathcal{A}$ given situational state S_t.

Regret. Before we formally define the learning objectives of the agent, we extend the agent state to account for randomized policies. We consider the notion of an algorithmic state Z_t introduced in Lu et al. (2021), which captures the algorithmic randomness at time t. An algorithm is a deterministic sequence $\{\mu_t \mid t = 1, 2, \ldots\}$ of functions, each mapping the pair $(\mathcal{H}_t, Z_t)$ to a policy. At each time step t, the algorithm samples a random algorithmic state Z_t and computes a policy $\pi_t = \mu_t(\mathcal{H}_t, Z_t)$. We also write $\pi_t \sim \mu_t(\mathcal{H}_t)$ when the randomness introduced by Z_t is clear in the context. The algorithm then samples actions $A_t \sim \pi_t(\cdot \mid S_t)$ at times t. For a policy π, we denote the average reward starting at state s as

$$\lambda_{\pi,\mathcal{E}}(s) = \liminf_{T\to\infty} \mathbb{E}_\pi \left[\frac{1}{T} \sum_{t=0}^{T-1} R_{t+1} \middle| \mathcal{E}, S_0 = s \right], \tag{1}$$

where the subscript of the expectation indicates that the reward sequence is realized by following policy π, and the subscript $\mathcal{E}$ emphasizes the dependence of the average reward on the environment $\mathcal{E}$. We also denote the optimal average reward as

$$\lambda_{*,\mathcal{E}}(s) = \sup_\pi \lambda_{\pi,\mathcal{E}}(s) \quad \forall s \in \mathcal{S}.$$

We consider weakly-communicating Markovian environments, the most general subclass of MDPs for which finite time regret bounds are plausible. This assumption also appears in Ouyang et al. (2017); Agrawal & Goyal (2013). We give a formal definition for MDPs with finite state spaces below.

Definition 2.1. (weakly-communicating MDP) *An MDP is weakly communicating if there exists a set of states, where each state in that set is accessible from every other state in that set under some deterministic stationary policy, along with a possibly empty set of states which is transient under every policy.*

We remark that the optimal average reward $\lambda_{*,\mathcal{E}}(\cdot)$ is state-independent under weakly-communicating MDPs. Thus, we override the notation $\lambda_{*,\mathcal{E}}$ to denote the optimal average reward $\lambda_{*,\mathcal{E}}(s)$ for all states $s \in \mathcal{S}$. For a policy π, we define its regret up to time T to be

$$\text{Regret}(T, \pi) \coloneqq \sum_{t=0}^{T-1} (\lambda_{*,\mathcal{E}} - R_{t+1}). \tag{2}$$

The regret itself is a random variable depending on the random environment $\mathcal{E}$, the algorithm's internal random sampling, and random transitions. We will measure agent performance in terms of regret and its expected value.

3 Continuing PSRL Algorithm

For episodic MDPs, the planning horizon is fixed ahead and known to the agent. The planning objective is often naturally the cumulative reward over the finite number of timesteps until the end of each episode. When the horizon is infinite, planning ahead becomes a challenge for the agent. One way to address this challenge is to set an effective finite planning horizon for the agent by maintaining a discount factor $\gamma \in [0, 1)$. Essentially, γ dictates the frequency with which the algorithm resamples an independent environment model used for planning. Given this discount factor, we divide the original infinite-horizon learning problem into *pseudo-episodes* with random lengths. Each pseudo-episode terminates when the algorithm resamples and computes a new policy. Concretely, at the beginning of timestep $t = 0, 1, \ldots,$ the agent samples a binary indicator X_t. If $X_t = 0$, the agent samples a new environment $\mathcal{E}$ based on the history $\mathcal{H}_t$ available at that time, and marks t as the start of a new pseudo-episode. It then computes a new policy π to follow in this pseudo-episode, and acts according to π. If $X_t = 1$, it continues the current pseudo-episode and follows the most recently computed policy π. When $X_t \sim \text{Bernoulli}(\gamma)$, one may interpret γ as the *survival probability* of a pseudo-episode at timestep t. We let E_k denote the set of timesteps in the k-th pseudo-episode, t_k denote the starting timestep of the k-th pseudo-episode, and $s_{k,1}$ denote the starting state of the k-th pseudo-episode.

Discounted value. At each timestep, the agent optimizes a discounted objective with the aforementioned discount factor γ. For each environment $\mathcal{E}$ and policy π, the γ-discounted value function $V^{\gamma}_{\pi,\mathcal{E}} \in \mathbb{R}^S$ of π in $\mathcal{E}$ is defined as

$$V^{\gamma}_{\pi,\mathcal{E}} \coloneqq \mathbb{E}\left[\sum_{h=0}^{H-1} P^h_\pi r_\pi \mid \mathcal{E}\right] = \mathbb{E}\left[\sum_{h=0}^{\infty} \gamma^h P^h_\pi r_\pi \mid \mathcal{E}\right], \tag{3}$$

where $P_{\pi ss'} = \sum_{a\in\mathcal{A}} \pi(a \mid s)\rho(s' \mid s, a)$ and $r_{\pi s} = \sum_{a\in\mathcal{A}} \pi(a \mid s) r_{as}$ for all $s, s' \in \mathcal{S}$ and $a \in \mathcal{A}$, and the expectation is over the random episode length H. Since a pseudo-episode terminates at time t when the independently sampled $X_t \sim \text{Bernoulli}(\gamma)$ takes value 0, its length H follows a binomial distribution with parameter γ. The second equality above is a direct consequence of this observation. A policy π is said to be optimal for the environment $\mathcal{E}$ if $V^{\gamma}_{\pi,\mathcal{E}} = \sup_{\pi'} V^{\gamma}_{\pi',\mathcal{E}}$. For an optimal policy π, we also write its value $V^{\gamma}_{*,\mathcal{E}}(s) \equiv V^{\gamma}_{\pi,\mathcal{E}}(s)$ as the optimal value. Furthermore, we denote an optimal policy with respect to $V^{\gamma}_{*,\mathcal{E}}$ for each $\mathcal{E}$ as $\pi^{\mathcal{E}}_{\gamma}$, which will be useful in the analysis. When γ is clear from the context, we omit the γ subscript to avoid cluttering. Note that $V^{\gamma}_{\pi,\mathcal{E}}$ satisfies the Bellman equation

$$V^{\gamma}_{\pi,\mathcal{E}} = r_\pi + \gamma P_\pi V^{\gamma}_{\pi,\mathcal{E}}. \tag{4}$$

Reward averaging time. We consider the notion of reward averaging times $\tau_{\pi,\mathcal{E}}$ of policies introduced in Dong et al. (2022) and derive regret bounds that depend on $\tau_{\pi,\mathcal{E}}$.

Definition 3.1. (reward averaging time) *The reward averaging time $\tau_{\pi,\mathcal{E}}$ of a policy π is the smallest value $\tau \in [0, \infty)$ such that*

$$\left|\mathbb{E}_\pi\left[\sum_{t=0}^{T-1} R_{t+1} \mid \mathcal{E}, S_0 = s\right] - T \cdot \lambda_{\pi,\mathcal{E}}(s)\right| \le \tau,$$

for all $T \ge 0$ and $s \in \mathcal{S}$.

When π^* is an optimal policy for $\mathcal{E}$, $\tau_{*,\mathcal{E}} \coloneqq \tau_{\pi^*,\mathcal{E}}$ is equivalent to the notion of span in Bartlett & Tewari (2009). We define Ω_* to be the set of all weakly communicating MDPs $\mathcal{E}$ and further make

the following assumption on the prior distribution $\mathbb{P}(\mathcal{E} \in \cdot)$. This assumption says that we focus on weakly communicating MDPs with bounded reward averaging time.

Assumption 3.2. *There exists $\tau < \infty$ such that the prior distribution over possible environments $\mathbb{P}(\mathcal{E} \in \cdot)$ satisfies*

$$\mathbb{P}\Big(\mathcal{E} \in \Omega_*,\ \tau_{*,\mathcal{E}} \leq \tau\Big) = 1.$$

Below, we restate an important lemma, Lemma 2 in Dong et al. (2022), that relates $\lambda_{\pi,\mathcal{E}}$, $\tau_{\pi,\mathcal{E}}$, and $V^{\gamma}_{\pi,\mathcal{E}}$.

Lemma 3.3. *For all π, $s \in \mathcal{S}$ and $\gamma \in [0,1)$,*

$$\left| V^{\gamma}_{\pi,\mathcal{E}}(s) - \frac{\lambda_{\pi,\mathcal{E}}(s)}{1-\gamma} \right| \leq \tau_{\pi,\mathcal{E}}.$$

We note again that for weakly communicating $\mathcal{E} \in \Omega_*$, the optimal average reward is state independent, i.e., $\lambda_{*,\mathcal{E}} = \lambda_{\pi_*,\mathcal{E}}(s)$ for all $s \in \mathcal{S}$. Thus, under Assumption 3.2, we have

$$\left| V^{\gamma}_{*,\mathcal{E}}(s) - V^{\gamma}_{*,\mathcal{E}}(s') \right| \leq 2\tau_{*,\mathcal{E}} \leq 2\tau \tag{5}$$

almost surely for all $s, s' \in \mathcal{S}$.

Discounted regret. Although the regret we eventually hope to analyze is defined by equation 2, we also consider a discounted version of the regret to aid our analysis. To analyze the performance of our algorithm over T timesteps, we set $K = \arg\max\{k : t_k \leq T\}$ to be the number of pseudo-episodes until time T. In our subsequent analysis, we adopt the convention that $t_{K+1} = T + 1$. To get a bound for general T, we can always fill in the rest of the timesteps to make a full pseudo-episode and the asymptotic rate stays the same. Moreover, it is easy to see that for all $\gamma \in [0,1)$, $\mathbb{E}[K] \leq (1-\gamma)T + 1$. Given a discount factor $\gamma \in [0,1)$, the γ-discounted regret up to time T is

$$\text{Regret}_{\gamma}(K, \pi) \coloneqq \sum_{k=1}^{K} \Delta_k, \tag{6}$$

where $\Delta_k = V^{\gamma}_{*,\mathcal{E}}(s_{k,1}) - V^{\gamma}_{\pi_k,\mathcal{E}}(s_{k,1})$ is the regret over pseudo-episode k, with $V^{\gamma}_{*,\mathcal{E}} = V^{\gamma}_{\pi^*,\mathcal{E}} = V^{\gamma}_{\pi^{\mathcal{E}},\mathcal{E}}$, $\pi_k \sim \mu_k(\mathcal{H}_{t_k})$, $A_t \sim \pi_k(\cdot \mid S_t)$, $S_{t+1} \sim \rho(\cdot|S_t, A_t)$, and $R_t = r(S_t, A_t, S_{t+1})$ for $t \in E_k$.

Empirical estimates. Finally, we define the empirical transition probabilities used by the algorithm. Let $N_t(s,a) = \sum_{\tau=1}^{t} \mathbb{1}\{(S_\tau, A_\tau) = (s,a)\}$ be the number of times action a has been sampled in state s up to timestep t. For every pair (s,a) with $N_{t_k}(s,a) > 0$, the empirical transition probabilities up to pseudo-episode k are

$$\hat{\rho}_k(s' \mid s, a) = \sum_{\ell=1}^{k-1} \sum_{t \in E_k} \frac{\mathbb{1}\{(S_t, A_t, S_{t+1}) = (s, a, s')\}}{N_{t_k}(s,a)} \tag{7}$$

for all $s' \in \mathcal{S}$. If the pair (s,a) has never been sampled before pseudo-episode k, we let $\hat{\rho}_k(s' \mid s,a) = 1$ for a random $s' \in \mathcal{S}$, and $\hat{\rho}_k(s'' \mid s,a) = 0$ for $s'' \in \mathcal{S} \setminus \{s'\}$. The corresponding matrix notation $\hat{P}^k$ are defined analogously.

3.1 The Algorithm

We present the Continuing Posterior Sampling for Reinforcement Learning algorithm in Algorithm 1, which extends PSRL Osband et al. (2013) to the infinite horizon setting with γ-discounted planning. The algorithm begins with a prior distribution over environments with actions $\mathcal{A}$ and state space $\mathcal{S}$. In addition, the algorithm takes an indicator $X_1 = 0$ and assumes a discount factor γ. At the start of each timestep t, if the indicator $X_t = 0$, Continuing PSRL samples an environment $\mathcal{E}_t = (\mathcal{A}, \mathcal{S}, \rho_t)$

from the posterior distribution conditioned on the history $\mathcal{H}_t$ available at that time, and mark t as the start of a new pseudo-episode. It then computes and follows the policy $\pi_t = \pi^{\mathcal{E}_t}$ at time t. Otherwise, if $X_t = 1$, it sticks to the policy $\pi_t = \pi_{t-1}$ at time t and adds step t to the current pseudo-episode. Then X_{t+1} is drawn from a Bernoulli distribution with parameter γ to be used in the next timestep.

Algorithm 1 Continuing PSRL (Continuing Posterior Sampling for Reinforcement Learning)

input Prior distribution f, discount factor γ, total learning time T
initialize $t = 1$, $k = 1$, $X_1 = 0$
for $t \leq T$ **do**
  if $X_t = 0$ **then**
    $t_k \leftarrow t$
    sample $\mathcal{E}^k \sim f(\cdot \mid \mathcal{H}_{t_k})$
    compute $\pi_k = \pi^{\mathcal{E}^k}$
    $k \leftarrow k + 1$
  end if
  sample and apply $A_t \sim \pi_k(\cdot \mid S_t)$
  observe R_{t+1} and S_{t+1}
  $t \leftarrow t + 1$
  sample $X_{t+1} \sim \text{Bernoulli}\,(\gamma)$
end for

Compared with the vanilla PSRL, Continuing PSRL simply adds an independent Bernoulli random number generator to determine when to resample. Although Continuing PSRL is not designed to be implemented in practice per se, we note that such resampling scheme brings both scalability and generalizability. For example, when the environment has an extremely large state or action space, e.g. Atari games Mnih et al. (2015), prior resampling methods relying on state-action visitation statistics Ouyang et al. (2017) require a huge look-up table, while the resampling method in Continuing PSRL can still be easily applied, with little computational overhead.

4 Main Results

We present our main results in this section. Theorem 4.1 establishes that PSRL with discounted planning satisfies a polynomial Bayesian regret bound for infinite-horizon tabular MDP environments. The bounds for the expected regret of Continuing PSRL are of order $\tilde{O}(\tau S\sqrt{AT})$, where $\tilde{O}$ omits logarithmic factors, matching the regret bound for TSDE in Ouyang et al. (2017), but achieved by a simple and elegant algorithm without additional episode termination criteria.

Theorem 4.1. *With $\gamma = 1 - \sqrt{SA/T}$, the regret of Algorithm 1 satisfies*

$$\mathbb{E}\left[\text{Regret}(T, \pi)\right] = \tilde{O}\left(\tau S\sqrt{AT}\right).$$

Note that our main theorem bounds the regret with respect to the optimal average reward, and thus has no dependence on the discount factor γ, which, as we emphasized, is only a design factor within the algorithm. For the purpose of the analysis, we utilize the γ-discounted regret defined in equation 6 and prove the following intermediate bound for the discounted regret.

Theorem 4.2. *Given T and $K = \arg\max\{k : t_k \leq T\}$,*

$$\mathbb{E}\left[\text{Regret}_\gamma(K, \pi)\right] \leq (4\tau + 2)S\sqrt{28AT\log(2SAT)} + \frac{SA}{1-\gamma}\log\left(\frac{1}{1-\gamma}\log\left(\frac{2T}{1-\gamma}\right)\right) + \frac{1}{2} + \frac{2}{1-\gamma}.$$

5 Analysis

As discussed in Osband et al. (2013), a key property of posterior sampling algorithms is that for any function g measurable with respect to the sigma algebra generated by the history, $\mathbb{E}[g(\mathcal{E})|\mathcal{H}_t] = \mathbb{E}[g(\mathcal{E}^t)|\mathcal{H}_t]$ if $\mathcal{E}^t$ is sampled from the posterior distribution at time t. Under our pseudo-episode construction, we state a stopping-time version of this property similar to the one in Ouyang et al. (2017).

Lemma 5.1. *If f is the distribution of $\mathcal{E}$, then for any $\sigma(\mathcal{H}_{t_k})$-measurable function g, we have*

$$\mathbb{E}[g(\mathcal{E})|\mathcal{H}_{t_k}] = \mathbb{E}[g(\mathcal{E}^k)|\mathcal{H}_{t_k}], \tag{8}$$

where the expectation is taken over f.

We include a proof in Appendix D.1 for completeness. We let $V^{\gamma}_{\pi_k,k} = V^{\gamma}_{\pi_k,\mathcal{E}^k}$ denote the value function of π_k, the policy employed by Continuing PSRL, under the sampled environment $\mathcal{E}^k$, and define

$$\tilde{\Delta}_k = V^{\gamma}_{\pi_k,k}(s_{k,1}) - V^{\gamma}_{\pi_k,\mathcal{E}}(s_{k,1})$$

as the difference in performance of π_k under $\mathcal{E}^k$ and the true environment $\mathcal{E}$. The next lemma allows us to evaluate regret in terms of $\tilde{\Delta}_k$, which we can analyze using the known sampled environment and our observation from the true environment, whereas it is typically hard to directly analyze Δ_k since we do not know what the optimal policy $\pi^* = \pi^{\mathcal{E}}$ is.

Lemma 5.2.

$$\mathbb{E}\left[\sum_{k=1}^{K} \Delta_k\right] = \mathbb{E}\left[\sum_{k=1}^{K} \tilde{\Delta}_k\right]. \tag{9}$$

Proof. The claim follows by applying Lemma 5.1 and a similar argument as that in Theorem 2 in Osband et al. (2013). □

We prove a value decomposition lemma that allows us to express the difference in values of π_k in the true environment $\mathcal{E}$ and sampled environment $\mathcal{E}^k$ in terms of a sum of Bellman errors over one stochastic pseudo-episode.

Lemma 5.3 (Value decomposition)**.** *For any environment $\hat{\mathcal{E}}$ and any policy π,*

$$\mathbb{E}\left[V^{\gamma}_{\pi,\mathcal{E}}(s_0) - V^{\gamma}_{\pi,\hat{\mathcal{E}}}(s_0) \mid \mathcal{E}, \hat{\mathcal{E}}\right] = \mathbb{E}\left[\sum_{t=0}^{\eta-1} \gamma \left\langle P_{\pi(s_t)s_t} - \hat{P}_{\pi(s_t)s_t}, V^{\gamma}_{\pi,\hat{\mathcal{E}}} \right\rangle \mid \mathcal{E}, \hat{\mathcal{E}}, \pi\right],$$

where η is the random length of a pseudo-episode, and the expectation is over the distribution of η, conditioned on the sampled state trajectory $s_0, s_1, \ldots$ drawn from following π in the environment $\mathcal{E}$.

We defer the proof to Appendix D.3. At the start of each pseudo-episode k, we consider a confidence set

$$\mathcal{M}_k = \left\{(\mathcal{A}, \mathcal{O}, \rho) : \left\|P_{as} - \hat{P}_{as}\right\|_1 \leq \sqrt{\frac{14S\log(2SAKt_k)}{\max\{N_{t_k}(s,a), 1\}}} \quad \forall (s,a)\right\}.$$

The following lemma bounds the probability that the true environment falls outside of the confidence set $\mathcal{M}_k$ for all $k \in [K]$.

Lemma 5.4. *For $k \in [K]$,*

$$\mathbb{P}\left(\mathcal{E} \notin \mathcal{M}_k\right) \leq \frac{1}{K}.$$

The proof follows standard concentration arguments, which we defer to Appendix D.3. We let

$$m = \frac{1}{1-\gamma} \log\left(\frac{2K}{1-\gamma}\right)$$

be a high probability upper bound for the length of each episode E_k, $k = 1, \dots, K$. Since $\tilde{\Delta}_k \leq \frac{1}{1-\gamma}$, we can decompose the regret as

$$\sum_{k=1}^{K} \tilde{\Delta}_k \leq \sum_{k=1}^{K} \tilde{\Delta}_k \mathbb{1}_{\{\mathcal{E}, \mathcal{E}^k \in \mathcal{M}_k\}} \cdot \mathbb{1}_{\{|E_k| \leq m\}} + \frac{1}{1-\gamma} \sum_{k=1}^{K} \left[\mathbb{1}_{\{|E_k| > m\}} + \mathbb{1}_{\{\mathcal{E} \notin \mathcal{M}_k\}} + \mathbb{1}_{\{\mathcal{E}^k \notin \mathcal{M}_k\}} \right].$$

Since $\mathcal{M}_k$ is $\sigma(\mathcal{H}_{t_k})$-measurable, by Lemma 5.1, $\mathbb{E}[\mathbb{1}_{\{\mathcal{E} \notin \mathcal{M}_k\}} \mid \mathcal{H}_{t_k}] = \mathbb{E}[\mathbb{1}_{\{\mathcal{E}^k \notin \mathcal{M}_k\}} \mid \mathcal{H}_{t_k}]$. Therefore,

$$\mathbb{E}\left[\sum_{k=1}^{K} \tilde{\Delta}_k\right] \leq \mathbb{E}\left[\sum_{k=1}^{K} \tilde{\Delta}_k \mathbb{1}_{\{\mathcal{E}, \mathcal{E}^k \in \mathcal{M}_k\}} \cdot \mathbb{1}_{\{|E_k| \leq m\}}\right] + \frac{1}{1-\gamma} \mathbb{E}\left[\sum_{k=1}^{K} \mathbb{1}_{\{|E_k| > m\}}\right] + \frac{2}{1-\gamma} \cdot \mathbb{E}\left[\sum_{k=1}^{K} \mathbb{1}_{\{\mathcal{E} \notin \mathcal{M}_k\}}\right]. \tag{10}$$

The third term can be bounded by $\frac{2}{1-\gamma}$ via Lemma 5.4. In what follows, we show how to bound the second and the first term.

Second term. For the second term, since $|E_k|$ follows a geometric distribution with parameter $1-\gamma$, applying the inequality $(1 + \frac{x}{n})^n \leq e^x$ by taking $n = \frac{1}{1-\gamma} \geq 1, x = -1$, we have

$$\frac{1}{1-\gamma} \mathbb{E}\left[\sum_{k=1}^{K} \mathbb{1}_{\{|E_k| > m\}}\right] = \frac{1}{1-\gamma} \sum_{k=1}^{K} \gamma^m = \frac{1}{1-\gamma} \sum_{k=1}^{K} \gamma^{\frac{1}{1-\gamma} \log(2K/(1-\gamma))} \leq \frac{1}{1-\gamma} \sum_{k=1}^{K} 2e^{-\log\left(\frac{2K}{1-\gamma}\right)} = \frac{1}{2}.$$

First term. It remains to bound the first term in equation 10. In Appendix A, we provide a detailed proof for the bound

$$\mathbb{E}\left[\sum_{k=1}^{K} \tilde{\Delta}_k \mathbb{1}_{\{\mathcal{E}, \mathcal{E}^k \in \mathcal{M}_k\}} \cdot \mathbb{1}_{\{|E_k| \leq m\}}\right] \leq 4\tau \cdot S\sqrt{28AT \log(2SAT)} + \frac{SA}{1-\gamma} \log\left(\frac{1}{1-\gamma} \log\left(\frac{2T}{1-\gamma}\right)\right).$$

Proof of Theorem 4.2. Combining the bound for each term in equation 10, we have

$$\mathbb{E}\left[\sum_{k=1}^{K} \tilde{\Delta}_k\right] \leq 4\tau S\sqrt{28AT \log(2SAT)} + \frac{SA}{1-\gamma} \log\left(\frac{1}{1-\gamma} \log\left(\frac{2T}{1-\gamma}\right)\right) + \frac{1}{2} + \frac{2}{1-\gamma}.$$

The claim follows from Lemma 5.2. □

5.1 Average Reward Regret

We have treated γ as constant so far, resulting in a regret bound in equation 6 that scales linearly with $\frac{1}{1-\gamma}$, ignoring logarithmic factors. However, the Continuing PSRL algorithm does not constrain γ to be a constant during the learning process, i.e. we are able to allow the discount factor to increase over time, accounting for the growing horizon. Specifically, at each time step t, we can consider a discount factor γ_t, and Continuing PSRL resamples a new environment with probability $1-\gamma_t$. If resampling happens, the agent switches to the optimal policy in the resampled environment maximizing the γ_t-discounted cumulative reward. Otherwise, the agent keeps following the previous policy. If γ_t increases with t, the agent is effectively planning over a longer horizon as interactions continue, and as Theorem 4.2 justifies, the agent's performance should always keep up with that of the optimal policy in the environment, regardless of the planning horizon. We detail in Appendix B a modified version of Continuing PSRL with a schedule of γ_t that attains the bound in Theorem 4.2.

Although Theorem 4.2 provides a sublinear upper bound for the discounted regret with a fixed discount factor γ, we are ultimately interested in the performance shortfall with respect to the optimal average-reward policy. To obtain a sublinear upper bound for the latter, we employ such time-dependent discount factors discussed above, which allows us to show Theorem 4.1, a regret bound for the Bayesian regret that does not depend on a discount factor, the proof of which is

relegated to Appendix C. If we assume the knowledge of T, optimizing for γ, the best rate in terms of T can be achieved by setting $1/(1-\gamma) = \sqrt{\frac{T}{SA}}$. If the total learning horizon T is unknown, we can utilize a classical doubling trick argument that is common in the design of online learning algorithms Zhang et al. (2020). The idea is to divide the learning horizon into time intervals of the form $[2^k, 2^{k+1})$ and set $1/(1-\gamma_t) = \sqrt{\frac{2^{k+1}}{SA}}$ for $t \in [2^k, 2^{k+1})$, $k \in \mathbb{N}$.

6 Simulations

We conduct two sets of simulations to empirically validate our proposed method. The first is a tabular RiverSwim environment introduced in (Osband et al., 2016a), and the second a RiverSwim environment with continuous features as observations. We modify both environments so that there is no episodic reset.

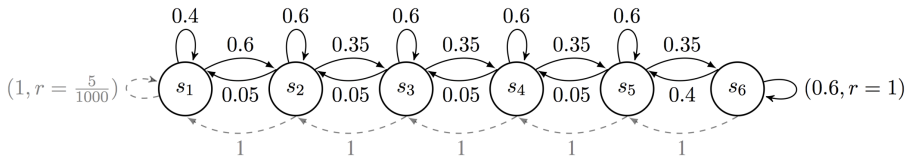

Figure 1: *Riverswim* - continuous and dotted arrows represent the MDP under the actions "right" and "left", respectively.

The tabular RiverSwim environment consists of six states arranged in a chain, as shown in Figure 1. The agent is placed at the far left state in the beginning. The agent receives a small reward of 0.005 for reaching the leftmost state, but the rightmost state contains a much larger reward of 1. The optimal policy is thus to always try to swim right.

We compare the performance of Continuing PSRL to TSDE (Ouyang et al., 2017) and DS-PSRL (Theocharous et al., 2018). Note that both baselines require planning with respect to the average reward directly. All agents start with a prior of Dirichlet and normal-gamma distributions over the transitions and rewards, respectively. We use pseudocount $n = 1$, $\alpha = 1/S$, and $\mu = \sigma^2 = 1$ for a diffuse uniform prior. Figure 2 (left) shows that Continuing PSRL is on par with TSDE despite the latter optimizing directly for average reward, and outperforms DS-PSRL by a large margin.

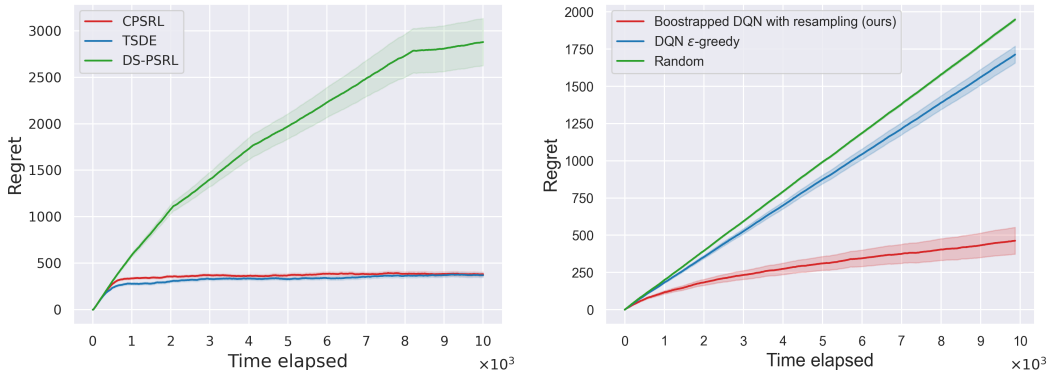

Figure 2: Regret curves aggregated across 50 Monte Carlo runs for all algorithms.

The second set of simulations aims to provide a concrete illustration of how our resampling technique can be extended and applied to complex scenarios. We consider a similar setup as RiverSwim, but with observations as raw pixel features, obtained via a feature mapping $\phi(s_t) = \mathbb{1}\{x \leq s_t\}$ in $[0,1]^N$. To extend our algorithm design to the continuous state space and model-free setting, we enhance the finite-horizon bootstrapped DQN with a similar resampling design as Continuing PSRL. The

idea is to resample a new ensemble index to use in each pseudo-episode if $X_t \sim \text{Bernoulli}(\gamma)$ equals zero. For this experiment, we choose a fixed $\gamma = 0.99$. The tabular baseline TSDE has to utilize visitation counts, and thus cannot be adapted to a continuous state space. DS-PSRL requires an oracle or subroutine that computes the optimal average reward, which is not readily available in their paper. Instead, we run a comparison against a uniformly random agent and a vanilla DQN agent with ϵ-greedy exploration, where we set $\epsilon = 0.1$. As shown in Figure 2 (right), Bootstrapped DQN with our random sampling schedule achieves sublinear regret, implying that the average regret tends to zero as T grows. In contrast, the vanilla DQN agent incurs linear regret, possibly due to its poor exploration ability.

7 Conclusion

We proposed a novel algorithm design extending PSRL to the setting where the environment does not have a reset schedule, and the agent has to plan over a possibly infinite horizon. We establish theoretically that our algorithm, Continuing PSRL, enjoys a regret upper bound that is close to the theoretical optimality. Notably, Continuing PSRL only relies on a single Bernoulli random number generator to resample the environment, as opposed to the complex episode-stopping schemes in prior works. Such design principle can be readily applied to general environments with large state spaces. Our simulations in both tabular and continuous RiverSwim environments demonstrate the effectiveness of our method. Moreover, Continuing PSRL also highlights the role of discount factor in agent design, as the discount factor is no longer considered as a part of the learning target, but mainly acts as a tool for the agent to dynamically adjust its planning horizon. As such, this work might provide an important step towards understanding discount factors, which have seen wide popularity in practical RL applications.

References

Shipra Agrawal and Navin Goyal. Further optimal regret bounds for Thompson sampling. In *International Conference on Artificial Intelligence and Statistics*, 2013.

Peter L. Bartlett and Ambuj Tewari. REGAL: A regularization based algorithm for reinforcement learning in weakly communicating MDPs. In *Conference on Uncertainty in Artificial Intelligence*, 2009.

Shi Dong, Benjamin Van Roy, and Zhengyuan Zhou. Simple agent, complex environment: Efficient reinforcement learning with agent states. *Journal of Machine Learning Research*, 23:1–54, 2022.

Vikranth Dwaracherla, Xiuyuan Lu, Morteza Ibrahimi, Ian Osband, Zheng Wen, and Benjamin Van Roy. Hypermodels for exploration. In *International Conference on Learning Representations*, 2020.

Vincent François-Lavet, Raphaël Fonteneau, and Damien Ernst. How to discount deep reinforcement learning: Towards new dynamic strategies. In *Advances in Neural Information Processing Systems*, 2015.

Tor Lattimore and Marcus Hutter. PAC bounds for discounted MDPs. In *International Conference on Algorithmic Learning Theory*, 2012.

Xiuyuan Lu, Benjamin Van Roy, Vikranth Dwaracherla, Morteza Ibrahimi, Ian Osband, and Zheng Wen. Reinforcement learning, bit by bit. *arXiv preprint arXiv:2103.04047*, 2021.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin A. Riedmiller, Andreas Fidjeland, Georg Ostrovski, Stig Petersen, Charlie Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518:529–533, 2015.

Abhishek Naik, Roshan Shariff, Niko Yasui, and Richard S. Sutton. Discounted reinforcement learning is not an optimization problem. *arXiv preprint arXiv:1910.02140*, 2019.

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

Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. Deep exploration via bootstrapped DQN. In *Advances in Neural Information Processing Systems*, 2016a.

Ian Osband, Benjamin Van Roy, and Zheng Wen. Generalization and exploration via randomized value functions. In *International Conference on International Conference on Machine Learning*, 2016b.

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*, 2017.

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

Dengwang Tang, Rahul Jain, Botao Hao, and Zheng Wen. Efficient online learning with offline datasets for infinite horizon mdps: A bayesian approach, 2024.

Georgios Theocharous, Zheng Wen, Yasin Abbasi Yadkori, and Nikos Vlassis. Scalar posterior sampling with applications. In *Advances in Neural Information Processing Systems*, 2018.

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

Yuanhao Wang, Kefan Dong, Xiaoyu Chen, and Liwei Wang. Q-learning with UCB exploration is sample efficient for infinite-horizon MDP. In *International Conference on Learning Representations*, 2020.

Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdú, and Marcelo J. Weinberger. Inequalities for the L1 deviation of the empirical distribution. In *Hewlett-Packard Labs Technical Reports*, 2003.

Junzi Zhang, Jongho Kim, Brendan O'Donoghue, and Stephen Boyd. Sample efficient reinforcement learning with REINFORCE. In *AAAI Conference on Artificial Intelligence*, 2020.

A Bounding the sum of confidence set widths

We are interested in bounding

$$\mathbb{E}\left[\sum_{k=1}^{K} \tilde{\Delta}_k \mathbb{1}_{\{\mathcal{E}, \mathcal{E}^k \in \mathcal{M}_k\}} \cdot \mathbb{1}_{\{|E_k| \leq m\}}\right].$$

Proof. For each k, we define the event

$$\mathcal{B}_k \coloneqq \left\{N_{t_{k+1}-1}(s,a) + 1 \leq 2N_{t_k}(s,a) \text{ for all } (s,a) \in \mathcal{S} \times \mathcal{A}\right\}.$$

Then $\mathcal{B}_k^c = \{N_{t_{k+1}-1}(s,a) \geq 2N_{t_k}(s,a) \text{ for some } (s,a) \in \mathcal{S} \times \mathcal{A}\}$.

Following a similar strategy as in Osband et al. (2013), we can write

$$\begin{aligned}
\mathbb{E}\left[\sum_{k=1}^{K} \tilde{\Delta}_k \mathbb{1}_{\{\mathcal{E}, \mathcal{E}^k \in \mathcal{M}_k\}} \cdot \mathbb{1}_{\{|E_k| \leq m\}}\right] &\leq \mathbb{E}\left[\sum_{k=1}^{K} \tilde{\Delta}_k \mathbb{1}_{\{\mathcal{E}, \mathcal{E}^k \in \mathcal{M}_k\}} \cdot \mathbb{1}_{\mathcal{B}_k}\right] + \mathbb{E}\left[\sum_{k=1}^{K} \tilde{\Delta}_k \mathbb{1}_{\{|E_k| \leq m\}} \cdot \mathbb{1}_{\mathcal{B}_k^c}\right] \\
&\leq \mathbb{E}\left[\sum_{k=1}^{K} \mathbb{E}\left[\tilde{\Delta}_k \mid \mathcal{E}, \mathcal{E}^k\right] \mathbb{1}_{\{\mathcal{E}, \mathcal{E}^k \in \mathcal{M}_k\}} \cdot \mathbb{1}_{\mathcal{B}_k}\right] + \mathbb{E}\left[\sum_{k=1}^{K} \tilde{\Delta}_k \mathbb{1}_{\{|E_k| \leq m\}} \cdot \mathbb{1}_{\mathcal{B}_k^c}\right].
\end{aligned}$$

The event $N_{t_{k+1}-1}(s,a) \geq 2N_{t_k}(s,a)$ can happen in at most $\log m$ episodes per state action pair under the event $\{|E_k| \leq m\}$. Thus, the second term can be bounded by

$$\mathbb{E}\left[\sum_{k=1}^{K} \tilde{\Delta}_k \mathbb{1}_{\{|E_k| \leq m\}} \cdot \mathbb{1}_{\mathcal{B}_k^c}\right] \leq \frac{SA}{1-\gamma} \log m.$$

We define $\{(s_{k,i}, a_{k,i})\}_{i=1}^{|E_k|}$ to be the trajectory followed by π_k in pseudo-episode k starting from state $s_{k,1} = s_{t_k}$, the state at the beginning of pseudo-episode k. By equation 5, $|\min_{s \in \mathcal{S}} V^{\gamma}_{\pi_k, \mathcal{E}^k}(s) - \max_{s \in \mathcal{S}} V^{\gamma}_{\pi_k, \mathcal{E}^k}(s)| \leq 2\tau$ for all $s \in \mathcal{S}$. Thus, we can bound the first term

$$\begin{aligned}
&\mathbb{E}\left[\sum_{k=1}^{K} \mathbb{E}\left[\tilde{\Delta}_k \mid \mathcal{E}, \mathcal{E}^k\right] \mathbb{1}_{\{\mathcal{E}, \mathcal{E}^k \in \mathcal{M}_k\}} \cdot \mathbb{1}_{\mathcal{B}_k}\right] \\
&\leq \mathbb{E} \sum_{k=1}^{K} \sum_{i=1}^{|E_k|} \mathbb{E}\left[\gamma \left|\langle P_{\pi_k(s_{k,i}) s_{k,i}} - P^k_{\pi_k(s_{k,i}) s_{k,i}}, V^{\gamma}_{\pi_k, \mathcal{E}^k}\rangle\right| \mid \mathcal{E}, \mathcal{E}^k\right] \cdot \mathbb{1}_{\{\mathcal{E}, \mathcal{E}^k \in \mathcal{M}_k\}} \cdot \mathbb{1}_{\mathcal{B}_k} \\
&= \mathbb{E} \sum_{k=1}^{K} \sum_{i=1}^{|E_k|} \mathbb{E}\Big[\gamma \left|\langle P_{\pi_k(s_{k,i}) s_{k,i}} - P^k_{\pi_k(s_{k,i}) s_{k,i}}, V^{\gamma}_{\pi_k, \mathcal{E}^k} - \min_{s \in \mathcal{S}} V^{\gamma}_{\pi_k, \mathcal{E}^k}(s) \cdot \mathbf{1}\rangle\right| \\
&\quad + \gamma \left|\langle P_{\pi_k(s_{k,i}) s_{k,i}} - P^k_{\pi_k(s_{k,i}) s_{k,i}}, \min_{s \in \mathcal{S}} V^{\gamma}_{\pi_k, \mathcal{E}^k}(s) \cdot \mathbf{1}\rangle\right| \mid \mathcal{E}, \mathcal{E}^k\Big] \cdot \mathbb{1}_{\{\mathcal{E}, \mathcal{E}^k \in \mathcal{M}_k\}} \cdot \mathbb{1}_{\mathcal{B}_k} \\
&= \mathbb{E} \sum_{k=1}^{K} \sum_{i=1}^{|E_k|} \mathbb{E}\Big[\gamma \left|\langle P_{\pi_k(s_{k,i}) s_{k,i}} - P^k_{\pi_k(s_{k,i}) s_{k,i}}, V^{\gamma}_{\pi_k, \mathcal{E}^k} - \min_{s \in \mathcal{S}} V^{\gamma}_{\pi_k, \mathcal{E}^k}(s) \cdot \mathbf{1}\rangle\right| \mid \mathcal{E}, \mathcal{E}^k\Big] \cdot \mathbb{1}_{\{\mathcal{E}, \mathcal{E}^k \in \mathcal{M}_k\}} \cdot \mathbb{1}_{\mathcal{B}_k} \\
&\leq \mathbb{E} \sum_{k=1}^{K} \sum_{i=1}^{|E_k|} \gamma \left\|P_{\pi_k(s_{k,i}) s_{k,i}} - P^k_{\pi_k(s_{k,i}) s_{k,i}}\right\|_1 \left\|V^{\gamma}_{\pi_k, \mathcal{E}^k} - \min_{s \in \mathcal{S}} V^{\gamma}_{\pi_k, \mathcal{E}^k}(s) \cdot \mathbf{1}\right\|_{\infty} \cdot \mathbb{1}_{\{\mathcal{E}, \mathcal{E}^k \in \mathcal{M}_k\}} \cdot \mathbb{1}_{\mathcal{B}_k} \\
&\leq \mathbb{E} \sum_{k=1}^{K} \sum_{i=1}^{|E_k|} \min\{4\tau\gamma\beta_k(s_{k,i}, a_{k,i}), 1\} \cdot \mathbb{1}_{\mathcal{B}_k}.
\end{aligned}$$

where the second equality follows since $P_{\pi_k(s_{k,i}) s_{k,i}}$ and $P^k_{\pi_k(s_{k,i}) s_{k,i}}$ are probability distributions, in the second-to-last inequality we apply Hölder's inequality, and in the last inequality we apply

Lemma 5.3. We proceed to bounding the first term. Recall that $\beta_k(s,a) = \sqrt{\frac{14S\log(2SAKt_k)}{\max\{N_{t_k}(s,a),1\}}}$, then

$$\sum_{k=1}^{K}\sum_{i=1}^{|E_k|}\min\{4\tau\gamma\beta_k(s_{k,i},a_{k,i}),1\}\cdot\mathbb{1}_{\mathcal{B}_k} \leq 4\tau\sum_{k=1}^{K}\sum_{i=1}^{|E_k|}\mathbb{1}_{\mathcal{B}_k}\sqrt{\frac{14S\log(2SAKt_k)}{\max\{1,N_{t_k}(s_{k,i},a_{k,i})\}}}.$$

Under the event $\mathcal{B}_k = \{N_{t_{k+1}-1}(s,a)+1 \leq 2N_{t_k}(s,a)\ \forall(s,a)\in\mathcal{S}\times\mathcal{A}\}$, for any $t\in E_k$, $N_t(s,a)+1 \leq N_{t_{k+1}-1}(s,a)+1 \leq 2N_{t_k}(s,a)$. Therefore,

$$\begin{aligned}
\sum_{k=1}^{K}\sum_{t\in E_k}\sqrt{\frac{\mathbb{1}_{\mathcal{B}_k}}{N_{t_k}(s_t,a_t)}} &\leq \sum_{k=1}^{K}\sum_{t\in E_k}\sqrt{\frac{2}{N_t(s_t,a_t)+1}} \\
&= \sqrt{2}\sum_{t=1}^{T}(N_t(s_t,a_t)+1)^{-1/2} \\
&\leq \sqrt{2}\sum_{s,a}\sum_{j=1}^{N_{T+1}(s,a)} j^{-1/2} \\
&\leq \sqrt{2}\sum_{s,a}\int_{x=0}^{N_{T+1}(s,a)} x^{-1/2}dx \\
&\leq \sqrt{2SA\sum_{s,a}N_{T+1}(s,a)} \\
&= \sqrt{2SAT}.
\end{aligned}$$

Since all rewards and transitions are absolutely constrained in $[0,1]$, our term of interest

$$\sum_{k=1}^{K}\sum_{i=1}^{|E_k|}\min\{4\tau\gamma\beta_k(s_{k,i},a_{k,i}),1\}\cdot\mathbb{1}_{\mathcal{B}_k} \leq 4\tau\cdot S\sqrt{28AT\log(2SAT)}.$$

□

B Continuing PSRL with γ-scheduling

Algorithm 2 Continuing PSRL with γ-scheduling

input Prior distribution f, discount factor schedule function `Schedule`, total learning time T

initialize $t=1$, $k=1$, $X_1=0$
for $t \leq T$ **do**
  if $X_t = 0$ **then**
    $t_k \leftarrow t$
    sample $\mathcal{E}^k \sim f(\cdot \mid \mathcal{H}_{t_k})$
    compute $\pi_k = \pi^{\mathcal{E}^k}$
    $k \leftarrow k+1$
  end if
  sample and apply $A_t \sim \pi_k(\cdot \mid S_t)$
  observe R_{t+1} and S_{t+1}
  $t \leftarrow t+1$
  $\gamma_t \leftarrow$ `Schedule`(t,T)
  sample $X_{t+1} \sim \text{Bernoulli}(\gamma_t)$
end for

C Proof of Theorem 4.1

Proof of Theorem 4.1. To better represent the dependence on K, set

$$\mathcal{R}(K,\pi) \coloneqq \text{Regret}(T,\pi),$$

where T is at the end of the K-th pseudo-episode. The expected regret can be written as

$$\mathbb{E}\left[\mathcal{R}(K,\pi)\right] = \mathbb{E}_\pi\left[\sum_{k=1}^{K}\sum_{t\in E_k}(\lambda_* - R_t)\right] = \mathbb{E}_\pi\left[\sum_{k=1}^{K}|E_k|\lambda_* - \sum_{k=1}^{K}\sum_{t\in E_k} R_t\right].$$

Adding and subtracting the optimal discounted value,

$$\mathbb{E}\left[\mathcal{R}(K,\pi)\right] = \underbrace{\mathbb{E}\left[\sum_{k=1}^{K}(|E_k|\lambda_* - V^{\gamma}_{*,\mathcal{E}}(s_{k,1}))\right]}_{(a)} + \underbrace{\mathbb{E}\left[\sum_{k=1}^{K}\left[V^{\gamma}_{*,\mathcal{E}}(s_{k,1}) - V^{\gamma}_{\pi_k,\mathcal{E}}(s_{k,1})\right]\right]}_{(b)}.$$

Since the length of each pseudo-episode is independent of the policy or the environment itself with mean $\frac{1}{1-\gamma}$, term (a) is the difference between the optimal average reward, weighted by the effective horizon, and the optimal discounted reward. By Lemma 3.3,

$$(a) = \mathbb{E}\left[\sum_{k=1}^{K}(\frac{1}{1-\gamma}\lambda_* - V^{\gamma}_{*,\mathcal{E}}(s_{k,1}))\right] = \mathbb{E}\left[\sum_{k=1}^{K}\left|\frac{1}{1-\gamma}\lambda_* - V^{\gamma}_{*,\mathcal{E}}(s_{k,1})\right|\right] \leq \tau\mathbb{E}[K].$$

The expectation (b) is the sum of differences between the optimal discounted value and the discounted value of the deployed policy over K pseudo-episodes. From equation 6 we can see that (b) is exactly $\mathbb{E}\left[\text{Regret}_\gamma(K,\pi)\right]$. By Theorem 4.2, we have that the regret can be bounded in terms of γ as

$$\mathbb{E}\left[\mathcal{R}(K,\pi)\right] \leq \tau\mathbb{E}[K] + \tilde{O}\left(4\tau S\sqrt{AT} + \frac{SA}{1-\gamma}\right) \leq (1-\gamma)T\tau + \tau + \tilde{O}\left(4\tau S\sqrt{AT} + \frac{SA}{1-\gamma}\right).$$

If we assume the knowledge of T, optimizing for γ, the best rate in terms of T can be achieved by setting $1/(1-\gamma) = \sqrt{\frac{T}{SA}}$, and the final bound becomes

$$\mathbb{E}[\text{Regret}(T,\pi)] = \mathbb{E}\left[\mathcal{R}(K,\pi)\right] \leq \tau\sqrt{SAT} + \tilde{O}\left(4\tau S\sqrt{AT}\right) = \tilde{O}\left(\tau S\sqrt{AT}\right).$$

If the total learning horizon T is unknown, we can utilize a classical doubling trick argument that is common in the design of online learning algorithms Zhang et al. (2020). The idea is to divide the learning horizon into time intervals of the form $[2^k, 2^{k+1})$ and set $1/(1-\gamma_t) = \sqrt{\frac{2^{k+1}}{SA}}$ for $t \in [2^k, 2^{k+1})$, $k \in \mathbb{N}$.

□

D Supporting Lemmas

D.1 Proof of Lemma 5.1

Proof. By definition, t_k is a stopping time, so it is $\sigma(\mathcal{H}_{t_k})$-measurable. Since $\mathcal{E}^k$ is sampled from the posterior distribution $\mathbb{P}(\mathcal{E} \in \cdot|\mathcal{H}_{t_k})$, $\mathcal{E}^k$ and $\mathbb{P}(\mathcal{E} \in \cdot|\mathcal{H}_{t_k})$ are also measurable with respect to $\sigma(\mathcal{H}_{t_k})$. Conditioning on $\mathcal{H}_{t_k}$, the only randomness in $g(\mathcal{E}^k)$ is the random sampling in the algorithm. The proof follows by integrating over $\mathbb{P}(\mathcal{E} \in \cdot|\mathcal{H}_{t_k})$. □

D.2 Proof of Lemma 5.3

Proof. Expanding the right hand side by the tower property, we get

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=0}^{\eta-1}\gamma\left\langle P_{\pi(s_t)s_t}-\hat{P}_{\pi(s_t)s_t}, V^{\gamma}_{\pi,\hat{\mathcal{E}}}\right\rangle \mid \mathcal{E},\hat{\mathcal{E}},\pi\right]\\
&=\sum_{H=1}^{\infty}\mathbb{E}\left[\sum_{t=0}^{H-1}\gamma\left\langle P_{\pi(s_t)s_t}-\hat{P}_{\pi(s_t)s_t}, V^{\gamma}_{\pi,\hat{\mathcal{E}}}\right\rangle \mid \mathcal{E},\hat{\mathcal{E}},\pi,\eta=H\right]\cdot\mathbb{P}(\eta=H)\\
&=\sum_{H=1}^{\infty}\mathbb{E}\left[\sum_{t=0}^{H-1}\gamma\left\langle P_{\pi(s_t)s_t}-\hat{P}_{\pi(s_t)s_t}, V^{\gamma}_{\pi,\hat{\mathcal{E}}}\right\rangle \mid \mathcal{E},\hat{\mathcal{E}},\pi\right]\cdot\gamma^{H-1}(1-\gamma)\\
&=\mathbb{E}\left[\sum_{t=0}^{\infty}\sum_{h=t}^{\infty}\gamma^{h}(1-\gamma)\cdot\gamma\left\langle P_{\pi(s_t)s_t}-\hat{P}_{\pi(s_t)s_t}, V^{\gamma}_{\pi,\hat{\mathcal{E}}}\right\rangle \mid \mathcal{E},\hat{\mathcal{E}},\pi\right]\\
&=\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^{t+1}\left\langle P_{\pi(s_t)s_t}-\hat{P}_{\pi(s_t)s_t}, V^{\gamma}_{\pi,\hat{\mathcal{E}}}\right\rangle \mid \mathcal{E},\hat{\mathcal{E}},\pi\right].
\end{aligned}$$

Now we consider the left hand side. By Bellman equations,

$$\begin{aligned}
V^{\gamma}_{\pi,\mathcal{E}}(s_0)-V^{\gamma}_{\pi,\hat{\mathcal{E}}}(s_0)&=r_{\pi s_0}+\gamma\langle P_{\pi(s_0)s_0},V^{\gamma}_{\pi,\mathcal{E}}\rangle-r_{\pi s_0}-\gamma\langle\hat{P}_{\pi(s_0)s_0},V^{\gamma}_{\pi,\hat{\mathcal{E}}}\rangle\\
&=\gamma\langle P_{\pi(s_0)s_0}-\hat{P}_{\pi(s_0)s_0},V^{\gamma}_{\pi,\hat{\mathcal{E}}}\rangle+\gamma\langle P_{\pi(s_0)s_0},V^{\gamma}_{\pi,\mathcal{E}}-V^{\gamma}_{\pi,\hat{\mathcal{E}}}\rangle\\
&=\gamma\langle P_{\pi(s_0)s_0}-\hat{P}_{\pi(s_0)s_0},V^{\gamma}_{\pi,\hat{\mathcal{E}}}\rangle+d_0+\gamma\left(V^{\gamma}_{\pi,\mathcal{E}}(s_1)-V^{\gamma}_{\pi,\hat{\mathcal{E}}}(s_1)\right),
\end{aligned}$$

where $d_i \coloneqq \gamma\langle P_{\pi(s_i)s_i},V^{\gamma}_{\pi,\mathcal{E}}-V^{\gamma}_{\pi,\hat{\mathcal{E}}}\rangle-\gamma\left(V^{\gamma}_{\pi,\mathcal{E}}(s_{i+1})-V^{\gamma}_{\pi,\hat{\mathcal{E}}}(s_{i+1})\right)$. Applying recursion, we have

$$V^{\gamma}_{\pi,\mathcal{E}}(s_0)-V^{\gamma}_{\pi,\hat{\mathcal{E}}}(s_0)=\sum_{t=0}^{\infty}\gamma^{t+1}\langle P_{\pi(s_t)s_t}-\hat{P}_{\pi(s_t)s_t},V^{\gamma}_{\pi,\mathcal{E}}\rangle+\sum_{t=1}^{\infty}d_t.$$

In state s_i under policy π, the expected value of $\gamma\left(V^{\gamma}_{\pi,\mathcal{E}}(s_{i+1})-V^{\gamma}_{\pi,\hat{\mathcal{E}}}(s_{i+1})\right)$ is exactly $\gamma\langle P_{\pi(s_i)s_i},V^{\gamma}_{\pi,\mathcal{E}}-V^{\gamma}_{\pi,\hat{\mathcal{E}}}\rangle$, so conditioning on the true environment $\mathcal{E}$ and the sampled environment $\hat{\mathcal{E}}$, the expectation of $\sum_{t=0}^{\infty}d_t$ is zero. Thus, the left hand side can be written as

$$\mathbb{E}\left[V^{\gamma}_{\pi,\mathcal{E}}(s_0)-V^{\gamma}_{\pi,\hat{\mathcal{E}}}(s_0)\mid\mathcal{E},\hat{\mathcal{E}}\right]=\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^{t+1}\langle P_{\pi(s_t)s_t}-\hat{P}_{\pi(s_t)s_t},V^{\gamma}_{\pi,\hat{\mathcal{E}}}\rangle\mid\mathcal{E},\hat{\mathcal{E}},\pi\right],$$

and our claim is proved. □

D.3 Proof of Lemma 5.4

Proof. By the L_1-deviation bound for empirical distributions in Weissman et al. (2003), when $N_{t_k}(s,a)>0$,

$$\begin{aligned}
\mathbb{P}\left(\left\|\hat{P}_{as}-P_{as}\right\|_1\geq\beta_k(s,a)|\mathcal{H}_{t_k}\right)&\leq(2^S-2)\exp\left(-\frac{N_{t_k}(s,a)\beta_k^2(s,a)}{2}\right)\\
&\leq 2^S\exp\left(-7S\log(2SAKt_k)\right)\leq 2^S\exp\big(-\log\big(4\cdot 2^S SAKt_k^2\big)\big)=\frac{1}{4SAKt_k^2},
\end{aligned}$$

where the second line is due to $\sum_{s'\in\mathcal{S}}\hat{P}_{as}(s')=\sum_{s'\in\mathcal{S}}P_{as}(s')=1$, and the third line follows from Hölder's inequality. When $N_{t_k}(s,a)=0$, the bounds trivially hold. Applying a union bound over

all possible values of $t_k = 1, \ldots, \infty$, we get

$$\mathbb{P}\left(\|\hat{P}_{as} - P_{as}\|_1 \geq \beta_k(s,a)\right) \leq \sum_{t=1}^{\infty} \frac{1}{4SAKt^2} = \frac{\pi^2}{24SAK} \leq \frac{1}{2SAK}.$$

We may conclude the proof with a union bound over all (s, a). $\square$