Online Learning for Stochastic Shortest Path Model via Posterior Sampling

Mehdi Jafarnia-Jahromi [1] **Liyu Chen** [1] **Rahul Jain** [1] **Haipeng Luo** [1]

Abstract

We consider the problem of online reinforcement learning for the Stochastic Shortest Path (SSP) problem modeled as an unknown MDP with an absorbing state. We propose `PSRL-SSP`, a simple posterior sampling-based reinforcement learning algorithm for the SSP problem. The algorithm operates in epochs. At the beginning of each epoch, a sample is drawn from the posterior distribution on the unknown model dynamics, and the optimal policy with respect to the drawn sample is followed during that epoch. An epoch completes if either the number of visits to the goal state in the current epoch exceeds that of the previous epoch, or the number of visits to any of the state-action pairs is doubled. We establish a Bayesian regret bound of $\tilde{\mathcal{O}}(B_\star S\sqrt{AK})$, where $B_\star$ is an upper bound on the expected cost of the optimal policy, S is the size of the state space, A is the size of the action space, and K is the number of episodes. The algorithm only requires the knowledge of the prior distribution, and has no hyper-parameters to tune. It is the first such posterior sampling algorithm and outperforms numerically previously proposed optimism-based algorithms.

1. Introduction

Stochastic Shortest Path (SSP) model considers the problem of an agent interacting with an environment to reach a predefined goal state while minimizing the cumulative expected cost. Unlike the finite-horizon and discounted Markov Decision Processes (MDPs) settings, in the SSP model, the horizon of interaction between the agent and the environment depends on the agent's actions, and can possibly be unbounded (if the goal is not reached). A wide variety of goal-oriented control and reinforcement learning (RL) problems such as navigation, game playing, etc. can be formulated as SSP problems. In the RL setting, where the SSP model is unknown, the agent interacts with the environment in K *episodes*. Each episode begins at a predefined initial state and ends when the agent reaches the goal (note that it might never reach the goal). We consider the *regret* minimization in the setting where the state and action spaces are finite, the cost function is known, but the transition kernel is unknown.

The agent has to balance the well-known trade-off between *exploration* and *exploitation*. A general way to balance the exploration-exploitation trade-off is to use the *Optimism in the Face of Uncertainty* (OFU) principle (Lai and Robbins, 1985). The idea is to construct a set of plausible models based on the available information, select the model associated with the minimum cost, and follow the optimal policy with respect to the selected model. This idea is widely used in the RL literature (e.g., (Jaksch et al., 2010; Jin et al., 2018; Wei et al., 2020; 2021; Tarbouriech et al., 2020; Rosenberg et al., 2020).)

An alternative fundamental idea to encourage exploration is to use Posterior Sampling (PS) (also known as Thompson Sampling) (Thompson, 1933). The idea is to maintain the posterior distribution on the unknown model parameters based on the available information and the prior distribution. PS algorithms usually proceed in *epochs*. In the beginning of an epoch, a model is sampled from the posterior. The actions during the epoch are then selected according to the optimal policy associated with the sampled model. PS algorithms have two main advantages over OFU-type algorithms. First, the prior knowledge of the environment can be incorporated through the prior distribution. Second, PS algorithms have shown superior numerical performance on multi-armed bandit problems (Scott, 2010; Chapelle and Li, 2011), and MDPs (Osband et al., 2013; Osband and Van Roy, 2017; Ouyang et al., 2017b).

The main difficulty in designing PS algorithms is the design of the epochs. In the basic setting of bandit problems, one can simply sample at every time step (Chapelle and Li, 2011). In finite-horizon MDPs, where the length of an episode is predetermined and fixed, the epochs and episodes coincide, i.e., the agent can sample from the posterior distribution at the beginning of each episode (Osband et al., 2013). However, in the general SSP model, where the length of each episode is not predetermined and can possibly be

[1]University of Southern California. Correspondence to: Mehdi Jafarnia-Jahromi <mjafarni@usc.edu>.

Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 2021.

unbounded, these natural choices for the epoch do not work. Indeed, the agent needs to switch policies during an episode if the current policy cannot reach the goal.

In this paper, we propose `PSRL-SSP`, the first PS-based RL algorithm for the SSP model. `PSRL-SSP` starts a new epoch based on two criteria. According to the first criterion, a new epoch starts if the number of episodes within the current epoch exceeds that of the previous epoch. The second criterion is triggered when the number of visits to any state-action pair is doubled during an epoch, similar to the one used by (Jaksch et al., 2010; Ouyang et al., 2017b; Rosenberg et al., 2020). Intuitively speaking, in the early stages of the interaction between the agent and the environment, the second criterion triggers more often. This criterion is responsible for switching policies during an episode if the current policy cannot reach the goal. In the later stages of the interaction, the first criterion triggers more often and encourages exploration. We prove a Bayesian regret bound of $\tilde{\mathcal{O}}(B_\star S\sqrt{AK})$, where S is the number of states, A is the number of actions, K is the number of episodes, and $B_\star$ is an upper bound on the expected cost of the optimal policy. This is similar to the regret bound of (Rosenberg et al., 2020) and has a gap of $\sqrt{S}$ with the minimax lower bound. We note that concurrent works of Tarbouriech et al. (2021b) and Cohen et al. (2021) have closed the gap via OFU algorithms and blackbox reduction to the finite-horizon, respectively. However, the goal of this paper is not to match the minimax regret bound, but rather to introduce the first PS algorithm that has near-optimal regret bound with superior numerical performance than OFU algorithms. This is verified with the experiments in Section 4. The $\sqrt{S}$ gap with the lower bound exists for the PS algorithms in the finite-horizon (Osband et al., 2013) and the infinite-horizon average-cost MDPs (Ouyang et al., 2017b) as well. Thus, it remains an open question whether it is possible to achieve the lower bound via PS algorithms in these settings.

1.1. Related Work

Posterior Sampling. The idea of PS algorithms dates back to the pioneering work of (Thompson, 1933). The algorithm was ignored for several decades until recently. In the past two decades, PS algorithms have successfully been developed for various settings including multi-armed bandits (e.g., (Scott, 2010; Chapelle and Li, 2011; Kaufmann et al., 2012; Agrawal and Goyal, 2012; 2013)), MDPs (e.g., (Strens, 2000; Osband et al., 2013; Fonteneau et al., 2013; Gopalan and Mannor, 2015; Osband and Van Roy, 2017; Kim, 2017; Ouyang et al., 2017b; Banjević and Kim, 2019)), Partially Observable MDPs (Jafarnia-Jahromi et al., 2021), and Linear Quadratic Control (e.g., (Abeille and Lazaric, 2017; Ouyang et al., 2017a)). The interested reader is referred to (Russo et al., 2017) and references therein for a more comprehensive literature review.

Online Learning in SSP. Another related line of work is online learning in the SSP model which was introduced by (Tarbouriech et al., 2020). They proposed an algorithm with $\tilde{\mathcal{O}}(K^{2/3})$ regret bound. Subsequent work of (Rosenberg et al., 2020) improved the regret bound to $\tilde{\mathcal{O}}(B_\star S\sqrt{AK})$. The concurrent works of (Cohen et al., 2021; Tarbouriech et al., 2021b) proved a minimax regret bound of $\tilde{\mathcal{O}}(B_\star\sqrt{SAK})$. However, none of these works propose a PS-type algorithm. We refer the interested reader to (Rosenberg and Mansour, 2020; Chen et al., 2020; Chen and Luo, 2021) for the SSP model with adversarial costs and (Tarbouriech et al., 2021a) for sample complexity of the SSP model with a generative model.

2. Preliminaries

A Stochastic Shortest Path (SSP) model is denoted by $\mathcal{M} = (\mathcal{S}, \mathcal{A}, c, \theta, s_{\text{init}}, g)$ where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $c : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the cost function, $s_{\text{init}} \in \mathcal{S}$ is the initial state, $g \notin \mathcal{S}$ is the goal state, and $\theta : \mathcal{S}^+ \times \mathcal{S} \times \mathcal{A} \to [0, 1]$ represents the transition kernel such that $\theta(s'|s, a) = \mathbb{P}(s'_t = s'|s_t = s, a_t = a)$ where $\mathcal{S}^+ = \mathcal{S} \cup \{g\}$ includes the goal state as well. Here $s_t \in \mathcal{S}$ and $a_t \in \mathcal{A}$ are the state and action at time $t = 1, 2, 3, \cdots$ and $s'_t \in \mathcal{S}^+$ is the subsequent state. We assume that the initial state s_{init} is a fixed and known state and $\mathcal{S}$ and $\mathcal{A}$ are finite sets with size S and A, respectively. A stationary policy is a deterministic map $\pi : \mathcal{S} \to \mathcal{A}$ that maps a state to an action. The *value function* (also called the *cost-to-go function*) associated with policy π is a function $V^\pi(\cdot; \theta) : \mathcal{S}^+ \to [0, \infty]$ given by $V^\pi(g; \theta) = 0$ and $V^\pi(s; \theta) := \mathbb{E}[\sum_{t=1}^{\tau_\pi(s)} c(s_t, \pi(s_t))|s_1 = s]$ for $s \in \mathcal{S}$, where $\tau_\pi(s)$ is the number of steps before reaching the goal state (a random variable) if the initial state is s and policy π is followed throughout the episode. Here, we use the notation $V^\pi(\cdot; \theta)$ to explicitly show the dependence of the value function on θ. Furthermore, the optimal value function can be defined as $V(s; \theta) = \min_\pi V^\pi(s; \theta)$. Policy π is called *proper* if the goal state is reached with probability 1, starting from any initial state and following π (i.e., $\max_s \tau_\pi(s) < \infty$ almost surely), otherwise it is called *improper*.

We consider the reinforcement learning problem of an agent interacting with an SSP model $\mathcal{M} = (\mathcal{S}, \mathcal{A}, c, \theta_*, s_{\text{init}}, g)$ whose transition kernel θ_* is randomly generated according to the prior distribution μ_1 at the beginning and is then fixed. We will focus on SSP models with transition kernels in the set $\Theta_{B_\star}$ with the following standard properties:

Assumption 1. *For all $\theta \in \Theta_{B_\star}$, the following holds: (1) there exists a proper policy, (2) for all improper policies π_θ, there exists a state $s \in \mathcal{S}$, such that $V^{\pi_\theta}(s; \theta) = \infty$, and (3) the optimal value function $V(\cdot; \theta)$ satisfies $\max_s V(s; \theta) \leq B_\star$.*

Bertsekas and Tsitsiklis (1991) prove that the first two conditions in Assumption 1 imply that for each $\theta \in \Theta_{B_\star}$, the optimal policy is stationary, deterministic, proper, and can be obtained by the minimizer of the *Bellman optimality equations* given by $V(s;\theta) =$

$$\min_a \Big\{ c(s,a) + \sum_{s' \in \mathcal{S}^+} \theta(s'|s,a) V(s';\theta) \Big\}, \quad \forall s \in \mathcal{S}. \quad (1)$$

Here, we assume that $\mathcal{S}$, $\mathcal{A}$, and the cost function c are known to the agent, however, the transition kernel θ_* is unknown. Moreover, we assume that the support of the prior distribution μ_1 is a subset of $\Theta_{B_\star}$.

The agent interacts with the environment in K episodes. Each episode starts from the initial state s_{init} and ends at the goal state g (note that the agent may never reach the goal). At each time t, the agent observes state s_t and takes action a_t. The environment then yields the next state $s'_t \sim \theta_*(\cdot|s_t, a_t)$. If the goal is reached (i.e., $s'_t = g$), then the current episode completes, a new episode starts, and $s_{t+1} = s_{\text{init}}$. If the goal is not reached (i.e., $s'_t \neq g$), then $s_{t+1} = s'_t$. The goal of the agent is to minimize the expected cumulative cost after K episodes, or equivalently, minimize the *Bayesian regret* defined as

$$R_K := \mathbb{E}\left[\sum_{t=1}^{T_K} c(s_t, a_t) - K V(s_{\text{init}}; \theta_*) \right],$$

where T_K is the total number of time steps before reaching the goal state for the Kth time, and $V(s_{\text{init}}; \theta_*)$ is the optimal value function from (1). Here, expectation is with respect to the prior distribution μ_1 for θ_*, the horizon T_K, the randomness in the state transitions, and the randomness of the algorithm. If the agent does not reach the goal state at any of the episodes (i.e., $T_K = \infty$), we define $R_K = \infty$.

3. A PSRL Algorithm for SSP Models

In this section, we propose the Posterior Sampling Reinforcement Learning (`PSRL-SSP`) algorithm (Algorithm 1) for the SSP model. The input of the algorithm is the prior distribution μ_1. At time t, the agent maintains the posterior distribution μ_t on the unknown parameter θ_* given by $\mu_t(\Theta) = \mathbb{P}(\theta_* \in \Theta | \mathcal{F}_t)$ for any set $\Theta \subseteq \Theta_{B_\star}$. Here $\mathcal{F}_t$ is the information available at time t (i.e., the sigma algebra generated by $s_1, a_1, \cdots, s_{t-1}, a_{t-1}, s_t$). Upon observing state s'_t by taking action a_t at state s_t, the posterior can be updated according to

$$\mu_{t+1}(d\theta) = \frac{\theta(s'_t|s_t, a_t)\mu_t(d\theta)}{\int \theta'(s'_t|s_t, a_t)\mu_t(d\theta')}. \quad (2)$$

The `PSRL-SSP` algorithm proceeds in epochs $\ell = 1, 2, 3, \cdots$. Let t_ℓ denote the start time of epoch ℓ. In the beginning of epoch ℓ, parameter θ_ℓ is sampled from the posterior distribution μ_{t_ℓ} and the actions within that epoch are chosen according to the optimal policy with respect to θ_ℓ. Each epoch ends if either of the two stopping criteria are satisfied. The first criterion is triggered if the number of visits to the goal state during the current epoch (denoted by K_ℓ) exceeds that of the previous epoch. This ensures that $K_\ell \leq K_{\ell-1} + 1$ for all ℓ. The second criterion is triggered if the number of visits to any of the state-action pairs is doubled compared to the beginning of the epoch. This guarantees that $n_t(s,a) \leq 2n_{t_\ell}(s,a)$ for all (s,a) where $n_t(s,a) = \sum_{\tau=1}^{t-1} \mathbf{1}_{\{s_\tau = s, a_\tau = a\}}$ denotes the number of visits to state-action pair (s,a) before time t.

The second stopping criterion is similar to that used by Jaksch et al. (2010); Rosenberg et al. (2020), and is one of the two stopping criteria used in the posterior sampling algorithm (`TSDE`) for the infinite-horizon average-cost MDPs (Ouyang et al., 2017b). This stopping criterion is crucial since it allows the algorithm to switch policies if the generated policy is improper and cannot reach the goal. We note that updating the policy only at the beginning of an episode (as done in the posterior sampling for finite-horizon MDPs (Osband et al., 2013)) does not work for SSP models, because if the generated policy in the beginning of the episode is improper, the goal is never reached and the regret is infinity.

The first stopping criterion is novel. A similar stopping criterion used in the posterior sampling for infinite-horizon MDPs (Ouyang et al., 2017b) is based on the length of the epochs, i.e., a new epoch starts if the length of the current epoch exceeds the length of the previous epoch. This leads to a bound of $\mathcal{O}(\sqrt{T_K})$ on the number of epochs which translates to a final regret bound of $\mathcal{O}(K^{2/3})$ in SSP models. However, our first stopping criterion allows us to bound the number of epochs by $\mathcal{O}(\sqrt{K})$ rather than $\mathcal{O}(\sqrt{T_K})$ (see Lemma 2). This is one of the key steps in avoiding dependency on $c_{\min}^{-1}$ (i.e., a lower bound on the cost function) in the main term of the regret and achieve a final regret bound of $\mathcal{O}(\sqrt{K})$.

Main Results. Our first result considers the case where the cost function is strictly positive for all state-action pairs. Subsequently, we extend the result to the general case. To facilitate the presentation of the results, we assume that $S \geq 2$, $A \geq 2$, and $K \geq S^2 A$. We first assume that

Assumption 2. *There exists $c_{\min} > 0$, such that $c(s,a) \geq c_{\min}$ for all state-action pairs (s,a).*

Theorem 1. *Suppose Assumptions 1 and 2 hold. Then, the regret of the `PSRL-SSP` algorithm is upper bounded as*

$$R_K = \mathcal{O}\left(B_\star S \sqrt{KA} L^2 + S^2 A \sqrt{\frac{{B_\star}^3}{c_{\min}}} L^2 \right),$$

Algorithm 1 `PSRL-SSP`

Input: μ_1

Initialization: $t \leftarrow 1, \ell \leftarrow 0, K_{-1} \leftarrow 0, t_0 \leftarrow 0, k_{t_0} \leftarrow 0$

for episodes $k = 1, 2, \cdots, K$ **do**
- $s_t \leftarrow s_{\text{init}}$
- **while** $s_t \neq g$ **do**
  - **if** $k - k_{t_\ell} > K_{\ell-1}$ *or* $n_t(s,a) > 2n_{t_\ell}(s,a)$ *for some* $(s,a) \in \mathcal{S} \times \mathcal{A}$ **then**
    - $K_\ell \leftarrow k - k_{t_\ell}$
    - $\ell \leftarrow \ell + 1$
    - $t_\ell \leftarrow t, k_{t_\ell} \leftarrow k$
    - Generate $\theta_\ell \sim \mu_{t_\ell}(\cdot)$ and compute $\pi_\ell(\cdot) = \pi^*(\cdot;\theta_\ell)$ according to (1)
  - Choose action $a_t = \pi_\ell(s_t)$ and observe $s'_t \sim \theta_*(\cdot|s_t, a_t)$
  - Update μ_{t+1} according to (2)
  - $s_{t+1} \leftarrow s'_t, t \leftarrow t+1$

where $L = \log(B_\star SAKc_{\min}^{-1})$.

A crucial point about the above result is that the dependency on $c_{\min}^{-1}$ is only in the lower order term. This allows us to extend the $\mathcal{O}(\sqrt{K})$ bound to the general case where Assumption 2 does not hold by using the perturbation technique of (Rosenberg et al., 2020) (see Theorem 2). Avoiding dependency on $c_{\min}^{-1}$ in the main term is achieved by using a Bernstein-type confidence set in the analysis inspired by (Rosenberg et al., 2020). We note that using a Hoeffding-type confidence set in the analysis as in Ouyang et al. (2017b) gives a regret bound of $\mathcal{O}(\sqrt{K/c_{\min}})$ which results in $\mathcal{O}(K^{2/3})$ regret bound if Assumption 2 is violated.

Theorem 2. *Suppose Assumption 1 holds. Running the* `PSRL-SSP` *algorithm with costs* $c_\epsilon(s,a) := \max\{c(s,a), \epsilon\}$ *for* $\epsilon = (S^2A/K)^{2/3}$ *yields* $R_K =$

$$\mathcal{O}\left(B_\star S\sqrt{KA}\tilde{L}^2 + (S^2A)^{\frac{2}{3}}K^{\frac{1}{3}}(B_\star^{\frac{3}{2}}\tilde{L}^2 + T_\star) + S^2AT_\star^{\frac{3}{2}}\tilde{L}^2\right),$$

where $\tilde{L} := \log(KB_\star T_\star SA)$.

Similar to the `Bernstein-SSP` algorithm (Rosenberg et al., 2020), for large enough K, the regret bounds scale as $\tilde{\mathcal{O}}(B_\star S\sqrt{KA})$, and have a gap of $\sqrt{S}$ with the lower bound.

4. Experiments

In this section, the performance of our `PSRL-SSP` algorithm is compared with existing OFU-type algorithms in the literature. Two environments are considered: RandomMDP and GridWorld. RandomMDP (Ouyang et al., 2017b; Wei et al., 2020) is a SSP with 8 states and 2 actions whose transition kernel and cost function are generated uniformly at random. GridWorld (Tarbouriech et al., 2020) is a 3×4 grid (total of 12 states including the goal state) and 4 actions (LEFT, RIGHT, UP, DOWN) with $c(s,a) = 1$ for any state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$. The agent starts from the initial state located at the top left corner of the grid, and ends in the goal state at the bottom right corner. At each time step, the agent attempts to move in one of the four directions. However, the attempt is successful only with probability 0.85. With probability 0.15, the agent takes any of the undesired directions uniformly at random. If the agent tries to move out of the boundary, the attempt will not be successful and it remains in the same position. We evaluate the frequentist regret of `PSRL-SSP` for a fixed environment (i.e., the environment is not sampled from a prior distribution). A Dirichlet prior with parameters $[0.1, \cdots, 0.1]$ is considered for the transition kernel.

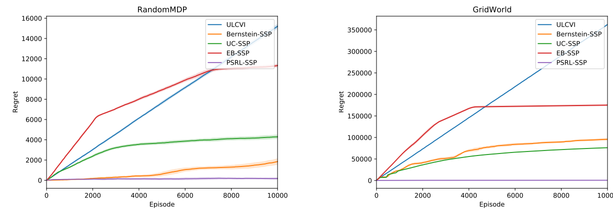

Figure 1. Cumulative regret of existing SSP algorithms on RandomMDP (left) and GridWorld (right) for $10,000$ episodes. The results are averaged over 10 runs and 95% confidence interval is shown with the shaded area. Our proposed `PSRL-SSP` algorithm outperforms all the existing algorithms considerably. The performance gap is even more significant in the more challenging GridWorld environment (right).

We compare the performance of our proposed `PSRL-SSP` against existing online learning algorithms for the SSP problem (`UC-SSP` (Tarbouriech et al., 2020), `Bernstein-SSP` (Rosenberg et al., 2020), `ULCVI` (Cohen et al., 2021), and `EB-SSP` (Tarbouriech et al., 2021b)).

Figure 1 shows that `PSRL-SSP` outperforms all the previously proposed algorithms for the SSP problem, significantly. In particular, it outperforms the recently proposed `ULCVI` (Cohen et al., 2021) and `EB-SSP` (Tarbouriech et al., 2021b) which match the theoretical lower bound. Our numerical evaluation reveals that the `ULCVI` algorithm does not show any evidence of learning even after 80,000 episodes (not shown here). The gap between the performance of `PSRL-SSP` and OFU algorithms is even more apparent in the GridWorld environment which is more challenging compared to RandomMDP. The poor performance of OFU algorithms ensures the necessity to consider PS algorithms in practice.

References

Marc Abeille and Alessandro Lazaric. Thompson sampling for linear-quadratic control problems. In *Artificial Intelligence and Statistics*, pages 1246–1254. PMLR, 2017.

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

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, pages 127–135. PMLR, 2013.

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.

Dragan Banjević and Michael Jong Kim. Thompson sampling for stochastic control: The continuous parameter case. *IEEE Transactions on Automatic Control*, 64(10): 4137–4152, 2019.

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

Dimitri P Bertsekas and John N Tsitsiklis. An analysis of stochastic shortest path problems. *Mathematics of Operations Research*, 16(3):580–595, 1991.

Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. *Advances in neural information processing systems*, 24:2249–2257, 2011.

Liyu Chen and Haipeng Luo. Finding the stochastic shortest path with low regret: The adversarial cost and unknown transition case. *arXiv preprint arXiv:2102.05284*, 2021.

Liyu Chen, Haipeng Luo, and Chen-Yu Wei. Minimax regret for stochastic shortest path with adversarial costs and known transition. *arXiv preprint arXiv:2012.04053*, 2020.

Alon Cohen, Yonathan Efroni, Yishay Mansour, and Aviv Rosenberg. Minimax regret for stochastic shortest path. *arXiv preprint arXiv:2103.13056*, 2021.

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.

Mehdi Jafarnia-Jahromi, Rahul Jain, and Ashutosh Nayyar. Online learning for unknown partially observable mdps. *arXiv preprint arXiv:2102.12661*, 2021.

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.

Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *International conference on algorithmic learning theory*, pages 199–213. Springer, 2012.

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

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

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

Ian Osband, Daniel Russo, and Benjamin Van Roy. (more) efficient reinforcement learning via posterior sampling. 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.

Aviv Rosenberg and Yishay Mansour. Stochastic shortest path with adversarially changing costs. *arXiv preprint arXiv:2006.11561*, 2020.

Aviv Rosenberg, Alon Cohen, Yishay Mansour, and Haim Kaplan. Near-optimal regret bounds for stochastic shortest path. In *International Conference on Machine Learning*, pages 8210–8219. PMLR, 2020.

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

Daniel Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, and Zheng Wen. A tutorial on thompson sampling. *arXiv preprint arXiv:1707.02038*, 2017.

Steven L Scott. A modern bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry*, 26(6):639–658, 2010.

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

Jean Tarbouriech, Evrard Garcelon, Michal Valko, Matteo Pirotta, and Alessandro Lazaric. No-regret exploration in goal-oriented reinforcement learning. In *International Conference on Machine Learning*, pages 9428–9437. PMLR, 2020.

Jean Tarbouriech, Matteo Pirotta, Michal Valko, and Alessandro Lazaric. Sample complexity bounds for stochastic shortest path with a generative model. In *Algorithmic Learning Theory*, pages 1157–1178. PMLR, 2021a.

Jean Tarbouriech, Runlong Zhou, Simon S Du, Matteo Pirotta, Michal Valko, and Alessandro Lazaric. Stochastic shortest path: Minimax, parameter-free and towards horizon-free regret. *arXiv preprint arXiv:2104.11186*, 2021b.

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.

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.

Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdu, and Marcelo J Weinberger. Inequalities for the l1 deviation of the empirical distribution. *Hewlett-Packard Labs, Tech. Rep*, 2003.

A. Theoretical Analysis

In this section, we prove Theorem 1. Proof of Theorem 2 can be found in the Appendix B.6.

A key property of posterior sampling is that conditioned on the information at time t, θ_* and θ_t have the same distribution if θ_t is sampled from the posterior distribution at time t (Osband et al., 2013; Russo and Van Roy, 2014). Since the PSRL-SSP algorithm samples θ_ℓ at the stopping time t_ℓ, we use the stopping time version of the posterior sampling property stated as follows.

Lemma 1 (Adapted from Lemma 2 of (Ouyang et al., 2017b))**.** *Let t_ℓ be a stopping time with respect to the filtration $(\mathcal{F}_t)_{t=1}^\infty$, and θ_ℓ be the sample drawn from the posterior distribution at time t_ℓ. Then, for any measurable function f and any $\mathcal{F}_{t_\ell}$-measurable random variable X, we have*

$$\mathbb{E}[f(\theta_\ell, X)|\mathcal{F}_{t_\ell}] = \mathbb{E}[f(\theta_*, X)|\mathcal{F}_{t_\ell}].$$

We now sketch the proof of Theorem 1. Let $0 < \delta < 1$ be a parameter to be chosen later. We distinguish between *known* and *unknown* state-action pairs. A state-action pair (s, a) is *known* if the number of visits to (s, a) is at least $\alpha \cdot \frac{B_\star S}{c_{\min}} \log \frac{B_\star SA}{\delta c_{\min}}$ for some large enough constant α (to be determined in Lemma 11), and *unknown* otherwise. We divide each epoch into *intervals*. The first interval starts at time $t = 1$. Each interval ends if any of the following conditions hold: (i) the total cost during the interval is at least $B_\star$; (ii) an unknown state-action pair is met; (iii) the goal state is reached; or (iv) the current epoch completes. The idea of introducing intervals is that after all state-action pairs are known, the cost accumulated during an interval is at least $B_\star$ (ignoring conditions (iii) and (iv)), which allows us to bound the number of intervals with the total cost divided by $B_\star$. Note that introducing intervals and distinguishing between known and unknown state-action pairs is only in the analysis and thus knowledge of $B_\star$ is not required.

Instead of bounding R_K, we bound R_M defined as

$$R_M := \mathbb{E}\left[\sum_{t=1}^{T_M} c(s_t, a_t) - KV(s_{\text{init}}; \theta_*)\right],$$

for any number of intervals M as long as K episodes are not completed. Here, T_M is the total time of the first M intervals. Let C_M denote the total cost of the algorithm after M intervals and define L_M as the number of epochs in the first M intervals. Observe that the number of times conditions (i), (ii), (iii), and (iv) trigger to start a new interval are bounded by $C_M/B_\star$, $\mathcal{O}(\frac{B_\star S^2 A}{c_{\min}} \log \frac{B_\star SA}{\delta c_{\min}})$, K, and L_M, respectively. Therefore, number of intervals can be bounded as

$$M \leq \frac{C_M}{B_\star} + K + L_M + \mathcal{O}(\frac{B_\star S^2 A}{c_{\min}} \log \frac{B_\star SA}{\delta c_{\min}}). \tag{3}$$

Moreover, since the cost function is lower bounded by $c_{\min}$, we have $c_{\min} T_M \leq C_M$. Our argument proceeds as follows.[1] We bound $R_M \lesssim B_\star S\sqrt{MA}$ which implies $\mathbb{E}[C_M] \lesssim K\mathbb{E}[V(s_{\text{init}}; \theta_*)] + B_\star S\sqrt{MA}$. From the definition of intervals and once all the state-action pairs are known, the cost accumulated within each interval is at least $B_\star$ (ignoring intervals that end when the epoch or episode ends). This allows us to bound the number of intervals M with $C_M/B_\star$ (or $\mathbb{E}[C_M]/B_\star$). Solving for $\mathbb{E}[C_M]$ in the quadratic inequality $\mathbb{E}[C_M] \lesssim K\mathbb{E}[V(s_{\text{init}}; \theta_*)] + B_\star S\sqrt{MA} \lesssim K\mathbb{E}[V(s_{\text{init}}; \theta_*)] + S\sqrt{\mathbb{E}[C_M]B_\star A}$ implies that $\mathbb{E}[C_M] \lesssim K\mathbb{E}[V(s_{\text{init}}; \theta_*)] + B_\star S\sqrt{AK}$. Since this bound holds for any number of M intervals as long as K episodes are not passed, it holds for $\mathbb{E}[C_K]$ as well. Moreover, since $c_{\min} > 0$, this implies that the K episodes eventually terminate and proves the final regret bound.

Bounding the Number of Epochs. Before proceeding with bounding R_M, we first prove that the number of epochs is bounded as $\mathcal{O}(\sqrt{KSA \log T_M})$. Recall that the length of the epochs is determined by two stopping criteria. If we ignore the second criterion for a moment, the first stopping criterion ensures that the number of episodes within each epoch grows at a linear rate which implies that the number of epochs is bounded by $\mathcal{O}(\sqrt{K})$. If we ignore the first stopping criterion for a moment, the second stopping criterion triggers at most $\mathcal{O}(SA \log T_M)$ times. The following lemma shows that the number of epochs remains of the same order even if these two criteria are considered simultaneously.

Lemma 2. *The number of epochs is bounded as $L_M \leq \sqrt{2SAK \log T_M} + SA \log T_M$.*

[1] Lower order terms are neglected.

We now provide the proof sketch for bounding R_M. With abuse of notation define $t_{L_M+1} := T_M + 1$. We can write

$$R_M := \mathbb{E}\left[\sum_{t=1}^{T_M} c(s_t, a_t) - KV(s_{\text{init}}; \theta_*)\right] = \mathbb{E}\left[\sum_{\ell=1}^{L_M} \sum_{t=t_\ell}^{t_{\ell+1}-1} c(s_t, a_t)\right] - K\mathbb{E}\left[V(s_{\text{init}}; \theta_*)\right]. \tag{4}$$

Note that within epoch ℓ, action a_t is taken according to the optimal policy with respect to θ_ℓ. Thus, with the Bellman equation we can write

$$c(s_t, a_t) = V(s_t; \theta_\ell) - \sum_{s'} \theta_\ell(s'|s_t, a_t) V(s'; \theta_\ell).$$

Substituting this and adding and subtracting $V(s_{t+1}; \theta_\ell)$ and $V(s'_t; \theta_\ell)$, decomposes R_M as

$$R_M = R_M^1 + R_M^2 + R_M^3,$$

where

$$\begin{aligned}
R_M^1 &:= \mathbb{E}\left[\sum_{\ell=1}^{L_M} \sum_{t=t_\ell}^{t_{\ell+1}-1} \left[V(s_t; \theta_\ell) - V(s_{t+1}; \theta_\ell)\right]\right] \\
R_M^2 &:= \mathbb{E}\left[\sum_{\ell=1}^{L_M} \sum_{t=t_\ell}^{t_{\ell+1}-1} \left[V(s_{t+1}; \theta_\ell) - V(s'_t; \theta_\ell)\right]\right] - K\mathbb{E}\left[V(s_{\text{init}}; \theta_*)\right] \\
R_M^3 &:= \mathbb{E}\left[\sum_{\ell=1}^{L_M} \sum_{t=t_\ell}^{t_{\ell+1}-1} \left[V(s'_t; \theta_\ell) - \sum_{s'} \theta_\ell(s'|s_t, a_t) V(s'; \theta_\ell)\right]\right].
\end{aligned}$$

We proceed by bounding these terms separately. Proof of these lemmas can be found in the supplementary material. R_M^1 is a telescopic sum and can be bounded by the following lemma.

Lemma 3. *The first term R_M^1 is bounded as $R_M^1 \leq B_\star \mathbb{E}[L_M]$.*

To bound R_M^2, recall that $s'_t \in \mathcal{S}^+$ is the next state of the environment after applying action a_t at state s_t, and that $s'_t = s_{t+1}$ for all time steps except the last time step of an episode (right before reaching the goal). In the last time step of an episode, $s'_t = g$ while $s_{t+1} = s_{\text{init}}$. This proves that the inner sum of R_M^2 can be written as $V(s_{\text{init}}; \theta_\ell) K_\ell$, where K_ℓ is the number of visits to the goal state during epoch ℓ. Using $K_\ell \leq K_{\ell-1} + 1$ and the property of posterior sampling completes the proof. This is formally stated in the following lemma.

Lemma 4. *The second term R_M^2 is bounded as $R_M^2 \leq B_\star \mathbb{E}[L_M]$.*

The rest of the proof proceeds to bound the third term R_M^3 which contributes to the dominant term of the final regret bound. The detailed proof can be found in Lemma 5. Here we provide the proof sketch. R_M^3 captures the difference between $V(\cdot; \theta_\ell)$ at the next state $s'_t \sim \theta_*(\cdot|s_t, a_t)$ and its expectation with respect to the sampled θ_ℓ. Applying the Hoeffding-type concentration bounds (Weissman et al., 2003), as used by (Ouyang et al., 2017b) yields a regret bound of $\mathcal{O}(K^{2/3})$ which is sub-optimal. To achieve the optimal dependency on K, we use a technique based on the Bernstein concentration bound inspired by the work of (Rosenberg et al., 2020). This requires a more careful analysis. Let $n_{t_\ell}(s, a, s')$ be the number of visits to state-action pair (s, a) followed by state s' before time t_ℓ. For a fixed state-action pair (s, a), define the Bernstein confidence set using the empirical transition probability $\hat{\theta}_\ell(s'|s, a) := \frac{n_{t_\ell}(s,a,s')}{n_{t_\ell}(s,a)}$ as

$$B_\ell(s, a) := \left\{\theta(\cdot|s, a) : |\theta(s'|s, a) - \hat{\theta}_\ell(s'|s, a)| \leq 4\sqrt{\hat{\theta}_\ell(s'|s, a) A_\ell(s, a)} + 28 A_\ell(s, a), \forall s' \in \mathcal{S}^+\right\}. \tag{5}$$

Here $A_\ell(s, a) := \frac{\log(SAn_\ell^+(s,a)/\delta)}{n_\ell^+(s,a)}$ and $n_\ell^+(s, a) := \max\{n_{t_\ell}(s, a), 1\}$. This confidence set is similar to the one used by Rosenberg et al. (2020) and contains the true transition probability $\theta_*(\cdot|s, a)$ with high probability (see Lemma 7). Note that $B_\ell(s, a)$ is $\mathcal{F}_{t_\ell}$-measurable which allows us to use the property of posterior sampling (Lemma 1) to conclude that $B_\ell(s, a)$

contains the sampled transition probability $\theta_\ell(\cdot|s,a)$ as well with high probability. With some algebraic manipulation, R_M^3 can be written as (with abuse of notation $\ell := \ell(t)$ is the epoch at time t)

$$R_M^3 = \mathbb{E}\left[\sum_{t=1}^{T_M}\sum_{s'\in\mathcal{S}}\left[\theta_*(s'|s_t,a_t)-\theta_\ell(s'|s_t,a_t)\right]\left(V(s';\theta_\ell)-\sum_{s''\in\mathcal{S}^+}\theta_*(s''|s_t,a_t)V(s'';\theta_\ell)\right)\right].$$

Under the event that both $\theta_*(\cdot|s_t,a_t)$ and $\theta_\ell(\cdot|s_t,a_t)$ belong to the confidence set $B_\ell(s_t,a_t)$, Bernstein bound can be applied to obtain

$$R_M^3 \approx \mathcal{O}\left(\mathbb{E}\left[\sum_{t=1}^{T_M}\sqrt{SA_\ell(s_t,a_t)\mathbb{V}_\ell(s_t,a_t)}\right]\right) = \mathcal{O}\left(\sum_{m=1}^{M}\mathbb{E}\left[\sum_{t=t_m}^{t_{m+1}-1}\sqrt{SA_\ell(s_t,a_t)\mathbb{V}_\ell(s_t,a_t)}\right]\right),$$

where t_m denotes the start time of interval m and $\mathbb{V}_\ell$ is the empirical variance defined as

$$\mathbb{V}_\ell(s_t,a_t) := \sum_{s'\in\mathcal{S}}\theta_*(s'|s_t,a_t)\left(V(s';\theta_\ell)-\sum_{s''\in\mathcal{S}^+}\theta_*(s''|s_t,a_t)V(s'';\theta_\ell)\right)^2. \tag{6}$$

Applying Cauchy Schwartz on the inner sum twice implies that

$$R_M^3 \approx \mathcal{O}\left(\sum_{m=1}^{M}\left(\sqrt{S\mathbb{E}\left[\sum_{t=t_m}^{t_{m+1}-1}A_\ell(s_t,a_t)\right]}\cdot\sqrt{\mathbb{E}\left[\sum_{t=t_m}^{t_{m+1}-1}\mathbb{V}_\ell(s_t,a_t)\right]}\right)\right)$$

Using the fact that all the state-action pairs (s_t,a_t) within an interval except possibly the first one are known, and that the cumulative cost within an interval is at most $2B_\star$, one can bound $\mathbb{E}\left[\sum_{t=t_m}^{t_{m+1}-1}\mathbb{V}_\ell(s_t,a_t)\right] = \mathcal{O}(B_\star^2)$ (see Lemma 10 for details). Applying Cauchy Schwartz again implies

$$R_M^3 \approx \mathcal{O}\left(B_\star\sqrt{MS\mathbb{E}\left[\sum_{t=1}^{T_M}A_\ell(s_t,a_t)\right]}\right) \approx \mathcal{O}\left(B_\star S\sqrt{MA}\right).$$

This argument is formally presented in the following lemma.

Lemma 5. *The third term R_M^3 can be bounded as*

$$R_M^3 \le 288B_\star S\sqrt{MA\log^2\frac{SA\mathbb{E}[T_M]}{\delta}} + 1632B_\star S^2A\log^2\frac{SA\mathbb{E}[T_M]}{\delta} + 4SB_\star\delta\mathbb{E}[L_M].$$

Detailed proofs of all lemmas and the theorem can be found in the appendix in the supplementary material.

B. Missing Proofs

B.1. Proof of Lemma 2

Lemma (restatement of Lemma 2)**.** The number of epochs is bounded as $L_M \le \sqrt{2SAK\log T_M} + SA\log T_M$.

Proof. Define macro epoch i with start time t_{u_i} given by $t_{u_1} = t_1$, and

$$t_{u_{i+1}} = \min\left\{t_\ell > t_{u_i} : n_{t_\ell}(s,a) > 2n_{t_\ell - 1}(s,a) \text{ for some } (s,a)\right\}, \qquad i = 2,3,\cdots.$$

A macro epoch starts when the second criterion of determining epoch length triggers. Let N_M be a random variable denoting the total number of macro epochs by the end of interval M and define $u_{N_M+1} := L_M + 1$.

Recall that K_ℓ is the number of visits to the goal state in epoch ℓ. Let $\tilde{K}_i := \sum_{\ell=u_i}^{u_{i+1}-1}K_\ell$ be the number of visits to the goal state in macro epoch i. By definition of macro epochs, all the epochs within a macro epoch except the last one are triggered by the first criterion, i.e., $K_\ell = K_{\ell-1}+1$ for $\ell = u_i,\cdots,u_{i+1}-2$. Thus,

$$\tilde{K}_i = \sum_{\ell=u_i}^{u_{i+1}-1}K_\ell = K_{u_{i+1}-1} + \sum_{j=1}^{u_{i+1}-u_i-1}(K_{u_i-1}+j) \ge \sum_{j=1}^{u_{i+1}-u_i-1} j = \frac{(u_{i+1}-u_i-1)(u_{i+1}-u_i)}{2}.$$

Solving for $u_{i+1} - u_i$ implies that $u_{i+1} - u_i \leq 1 + \sqrt{2\tilde{K}_i}$. We can write

$$L_M = u_{N_M+1} - 1 = \sum_{i=1}^{N_M} (u_{i+1} - u_i) \leq \sum_{i=1}^{N_M} \left(1 + \sqrt{2\tilde{K}_i}\right) = N_M + \sum_{i=1}^{N_M} \sqrt{2\tilde{K}_i}$$
$$\leq N_M + \sqrt{2N_M \sum_{i=1}^{N_M} \tilde{K}_i} = N_M + \sqrt{2N_M K},$$

where the second inequality follows from Cauchy-Schwartz. It suffices to show that the number of macro epochs is bounded as $N_M \leq 1 + SA \log T_M$. Let $\mathcal{T}_{s,a}$ be the set of all time steps at which the second criterion is triggered for state-action pair (s, a), i.e.,

$$\mathcal{T}_{s,a} := \left\{t_\ell \leq T_M : n_{t_\ell}(s, a) > 2n_{t_{\ell-1}}(s, a)\right\}.$$

We claim that $|\mathcal{T}_{s,a}| \leq \log n_{T_M+1}(s, a)$. To see this, assume by contradiction that $|\mathcal{T}_{s,a}| \geq 1 + \log n_{T_M+1}(s, a)$, then

$$n_{t_{L_M}}(s, a) = \prod_{t_\ell \leq T_M, n_{t_{\ell-1}}(s,a) \geq 1} \frac{n_{t_\ell}(s, a)}{n_{t_{\ell-1}}(s, a)} \geq \prod_{t_\ell \in \mathcal{T}_{s,a}, n_{t_{\ell-1}}(s,a) \geq 1} \frac{n_{t_\ell}(s, a)}{n_{t_{\ell-1}}(s, a)}$$
$$> 2^{|\mathcal{T}_{s,a}|-1} \geq n_{T_M+1}(s, a),$$

which is a contradiction. Thus, $|\mathcal{T}_{s,a}| \leq \log n_{T_M+1}(s, a)$ for all (s, a). In the above argument, the first inequality is by the fact that $n_t(s, a)$ is non-decreasing in t, and the second inequality is by the definition of $\mathcal{T}_{s,a}$. Now, we can write

$$N_M = 1 + \sum_{s,a} |\mathcal{T}_{s,a}| \leq 1 + \sum_{s,a} \log n_{T_M+1}(s, a)$$
$$\leq 1 + SA \log \frac{\sum_{s,a} n_{T_M+1}(s, a)}{SA} = 1 + SA \log \frac{T_M}{SA} \leq SA \log T_M,$$

where the second inequality follows from Jensen's inequality. □

B.2. Proof of Lemma 3

Lemma (restatement of Lemma 3)**.** The first term R_M^1 is bounded as $R_M^1 \leq B_\star \mathbb{E}[L_M]$.

Proof. Recall

$$R_M^1 = \mathbb{E}\left[\sum_{\ell=1}^{L_M} \sum_{t=t_\ell}^{t_{\ell+1}-1} [V(s_t; \theta_\ell) - V(s_{t+1}; \theta_\ell)]\right]$$

Observe that the inner sum is a telescopic sum, thus

$$R_M^1 = \mathbb{E}\left[\sum_{\ell=1}^{L_M} \left[V(s_{t_\ell}; \theta_\ell) - V(s_{t_{\ell+1}}; \theta_\ell)\right]\right] \leq B_\star \mathbb{E}[L_M],$$

where the inequality is by Assumption 1. □

B.3. Proof of Lemma 4

Lemma (restatement of Lemma 4)**.** The second term R_M^2 is bounded as $R_M^2 \leq B_\star \mathbb{E}[L_M]$.

Proof. Recall that K_ℓ is the number of times the goal state is reached during epoch ℓ. By definition, the only time steps that

$s'_t \neq s_{t+1}$ is right before reaching the goal. Thus, with $V(g;\theta_\ell)=0$, we can write

$$\begin{aligned}
R_M^2 &= \mathbb{E}\left[\sum_{\ell=1}^{L_M}\sum_{t=t_\ell}^{t_{\ell+1}-1}\left[V(s_{t+1};\theta_\ell)-V(s'_t;\theta_\ell)\right]\right]-K\mathbb{E}\left[V(s_{\text{init}};\theta_*)\right] \\
&= \mathbb{E}\left[\sum_{\ell=1}^{L_M}V(s_{\text{init}};\theta_\ell)K_\ell\right]-K\mathbb{E}\left[V(s_{\text{init}};\theta_*)\right] \\
&= \sum_{\ell=1}^{\infty}\mathbb{E}\left[\mathbf{1}_{\{m(t_\ell)\le M\}}V(s_{\text{init}};\theta_\ell)K_\ell\right]-K\mathbb{E}\left[V(s_{\text{init}};\theta_*)\right],
\end{aligned}$$

where the last step is by Monotone Convergence Theorem. Here $m(t_\ell)$ is the interval at time t_ℓ. Note that from the first stopping criterion of the algorithm we have $K_\ell \le K_{\ell-1}+1$ for all ℓ. Thus, each term in the summation can be bounded as

$$\mathbb{E}\left[\mathbf{1}_{\{m(t_\ell)\le M\}}V(s_{\text{init}};\theta_\ell)K_\ell\right]\le\mathbb{E}\left[\mathbf{1}_{\{m(t_\ell)\le M\}}V(s_{\text{init}};\theta_\ell)(K_{\ell-1}+1)\right].$$

$\mathbf{1}_{\{m(t_\ell)\le M\}}(K_{\ell-1}+1)$ is $\mathcal{F}_{t_\ell}$ measurable. Therefore, applying the property of posterior sampling (Lemma 1) implies

$$\mathbb{E}\left[\mathbf{1}_{\{m(t_\ell)\le M\}}V(s_{\text{init}};\theta_\ell)(K_{\ell-1}+1)\right]=\mathbb{E}\left[\mathbf{1}_{\{m(t_\ell)\le M\}}V(s_{\text{init}};\theta_*)(K_{\ell-1}+1)\right]$$

Substituting this into R_M^2, we obtain

$$\begin{aligned}
R_M^2 &\le \sum_{\ell=1}^{\infty}\mathbb{E}\left[\mathbf{1}_{\{m(t_\ell)\le M\}}V(s_{\text{init}};\theta_*)(K_{\ell-1}+1)\right]-K\mathbb{E}\left[V(s_{\text{init}};\theta_*)\right] \\
&= \mathbb{E}\left[\sum_{\ell=1}^{L_M}V(s_{\text{init}};\theta_*)(K_{\ell-1}+1)\right]-K\mathbb{E}\left[V(s_{\text{init}};\theta_*)\right] \\
&= \mathbb{E}\left[V(s_{\text{init}};\theta_*)\left(\sum_{\ell=1}^{L_M}K_{\ell-1}-K\right)\right]+\mathbb{E}\left[V(s_{\text{init}};\theta_*)L_M\right]\le B_\star\mathbb{E}[L_M].
\end{aligned}$$

In the last inequality we have used the fact that $0\le V(s_{\text{init}};\theta_*)\le B_\star$ and $\sum_{\ell=1}^{L_M}K_{\ell-1}\le K$. □

B.4. Proof of Lemma 5

Lemma (restatement of Lemma 5)**.** The third term R_M^3 can be bounded as

$$R_M^3\le 288B_\star S\sqrt{MA\log^2\frac{SA\mathbb{E}[T_M]}{\delta}}+1632B_\star S^2A\log^2\frac{SA\mathbb{E}[T_M]}{\delta}+4SB_\star\delta\mathbb{E}[L_M].$$

Proof. With abuse of notation let $\ell:=\ell(t)$ denote the epoch at time t and $m(t)$ be the interval at time t. We can write

$$\begin{aligned}
R_M^3 &= \mathbb{E}\left[\sum_{t=1}^{T_M}\left[V(s'_t;\theta_\ell)-\sum_{s'}\theta_\ell(s'|s_t,a_t)V(s';\theta_\ell)\right]\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{\infty}\mathbf{1}_{\{m(t)\le M\}}\left[V(s'_t;\theta_\ell)-\sum_{s'}\theta_\ell(s'|s_t,a_t)V(s';\theta_\ell)\right]\right] \\
&= \sum_{t=1}^{\infty}\mathbb{E}\left[\mathbf{1}_{\{m(t)\le M\}}\mathbb{E}\left[V(s'_t;\theta_\ell)-\sum_{s'}\theta_\ell(s'|s_t,a_t)V(s';\theta_\ell)\Big|\mathcal{F}_t,\theta_*,\theta_\ell\right]\right].
\end{aligned}$$

The last equality follows from Dominated Convergence Theorem, tower property of conditional expectation, and that $\mathbf{1}_{\{m(t)\le M\}}$ is measurable with respect to $\mathcal{F}_t$. Note that conditioned on $\mathcal{F}_t$, θ_* and θ_ℓ, the only random variable in the inner expectation is s'_t. Thus, $\mathbb{E}[V(s'_t;\theta_\ell)|\mathcal{F}_t,\theta_*,\theta_\ell]=\sum_{s'}\theta_*(s'|s_t,a_t)V(s';\theta_\ell)$. Using Dominated Convergence Theorem again implies that

$$
\begin{aligned}
R_M^3 &= \mathbb{E}\left[\sum_{t=1}^{T_M}\sum_{s'\in\mathcal{S}}\left[\theta_*(s'|s_t,a_t)-\theta_\ell(s'|s_t,a_t)\right]V(s';\theta_\ell)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T_M}\sum_{s'\in\mathcal{S}}\left[\theta_*(s'|s_t,a_t)-\theta_\ell(s'|s_t,a_t)\right]\left(V(s';\theta_\ell)-\sum_{s''\in\mathcal{S}^+}\theta_*(s''|s_t,a_t)V(s'';\theta_\ell)\right)\right], \qquad (7)
\end{aligned}
$$

where the last equality is due to the fact that $\theta_*(\cdot|s_t,a_t)$ and $\theta_\ell(\cdot|s_t,a_t)$ are probability distributions and that $\sum_{s''\in\mathcal{S}^+}\theta_*(s''|s_t,a_t)V(s'';\theta_\ell)$ is independent of s'.

Recall the Bernstein confidence set $B_\ell(s,a)$ defined in (5) and let $\Omega_{s,a}^\ell$ be the event that both $\theta_*(\cdot|s,a)$ and $\theta_\ell(\cdot|s,a)$ are in $B_\ell(s,a)$. If $\Omega_{s,a}^\ell$ holds, then the difference between $\theta_*(\cdot|s,a)$ and $\theta_\ell(\cdot|s,a)$ can be bounded by the following lemma.

Lemma 6. *Denote $A_\ell(s,a)=\frac{\log(SAn_\ell^+(s,a)/\delta)}{n_\ell^+(s,a)}$. If $\Omega_{s,a}^\ell$ holds, then*

$$
|\theta_*(s'|s,a)-\theta_\ell(s'|s,a)|\le 8\sqrt{\theta_*(s'|s,a)A_\ell(s,a)}+136A_\ell(s,a).
$$

Proof. Since $\Omega_{s,a}^\ell$ holds, by (5) we have that

$$
\hat\theta_\ell(s'|s,a)-\theta_*(s'|s,a)\le 4\sqrt{\hat\theta_\ell(s'|s,a)A_\ell(s,a)}+28A_\ell(s,a).
$$

Using the primary inequality that $x^2\le ax+b$ implies $x\le a+\sqrt{b}$ with $x=\sqrt{\hat\theta_\ell(s'|s,a)}$, $a=4\sqrt{A_\ell(s,a)}$, and $b=\theta_*(s'|s,a)+28A_\ell(s,a)$, we obtain

$$
\sqrt{\hat\theta_\ell(s'|s,a)}\le 4\sqrt{A_\ell(s,a)}+\sqrt{\theta_*(s'|s,a)+28A_\ell(s,a)}\le\sqrt{\theta_*(s'|s,a)}+10\sqrt{A_\ell(s,a)},
$$

where the last inequality is by sub-linearity of the square root. Substituting this bound into (5) yields

$$
|\theta_*(s'|s,a)-\hat\theta_\ell(s'|s,a)|\le 4\sqrt{\theta_*(s'|s,a)A_\ell(s,a)}+68A_\ell(s,a).
$$

Similarly,

$$
|\theta_\ell(s'|s,a)-\hat\theta_\ell(s'|s,a)|\le 4\sqrt{\theta_*(s'|s,a)A_\ell(s,a)}+68A_\ell(s,a).
$$

Using the triangle inequality completes the proof. □

Note that if either of $\theta_*(\cdot|s_t,a_t)$ or $\theta_\ell(\cdot|s_t,a_t)$ are not in $B_\ell(s_t,a_t)$, then the inner term of (7) can be bounded by $2SB_\star$ (note that $|\mathcal{S}^+|\le 2S$ and $V(\cdot;\theta_\ell)\le B_\star$). Thus, applying Lemma 6 implies that

$$
\begin{aligned}
&\sum_{s'\in\mathcal{S}}\left[\theta_*(s'|s_t,a_t)-\theta_\ell(s'|s_t,a_t)\right]\left(V(s';\theta_\ell)-\sum_{s''\in\mathcal{S}^+}\theta_*(s''|s_t,a_t)V(s'';\theta_\ell)\right) \\
&\le 8\sum_{s'\in\mathcal{S}}\sqrt{A_\ell(s_t,a_t)\theta_*(s'|s_t,a_t)\left(V(s';\theta_\ell)-\sum_{s''\in\mathcal{S}^+}\theta_*(s''|s_t,a_t)V(s'';\theta_\ell)\right)^2}\mathbf{1}_{\Omega^\ell_{s_t,a_t}} \\
&\quad+136\sum_{s'\in\mathcal{S}}A_\ell(s_t,a_t)\left|V(s';\theta_\ell)-\sum_{s''\in\mathcal{S}^+}\theta_*(s''|s_t,a_t)V(s'';\theta_\ell)\right|\mathbf{1}_{\Omega^\ell_{s_t,a_t}} \\
&\quad+2SB_\star\left(\mathbf{1}_{\{\theta_*(\cdot|s_t,a_t)\notin B_\ell(s_t,a_t)\}}+\mathbf{1}_{\{\theta_\ell(\cdot|s_t,a_t)\notin B_\ell(s_t,a_t)\}}\right) \\
&\le 16\sqrt{SA_\ell(s_t,a_t)\mathbb{V}_\ell(s_t,a_t)}\mathbf{1}_{\Omega^\ell_{s_t,a_t}}+272SB_\star A_\ell(s_t,a_t)\mathbf{1}_{\Omega^\ell_{s_t,a_t}} \\
&\quad+2SB_\star\big(\mathbf{1}_{\{\theta_*(\cdot|s_t,a_t)\notin B_\ell(s_t,a_t)\}}+\mathbf{1}_{\{\theta_\ell(\cdot|s_t,a_t)\notin B_\ell(s_t,a_t)\}}\big).
\end{aligned}
$$

where $A_\ell(s,a) = \frac{\log(SAn_\ell^+(s,a)/\delta)}{n_\ell^+(s,a)}$ and $\mathbb{V}_\ell(s,a)$ is defined in (6). Here the last inequality follows from Cauchy-Schwartz, $|\mathcal{S}^+| \leq 2S$, $V(\cdot;\theta_\ell) \leq B_\star$ and the definition of $\mathbb{V}_\ell$. Substituting this into (7) yields

$$R_M^3 \leq 16\sqrt{S}\mathbb{E}\left[\sum_{t=1}^{T_M} \sqrt{A_\ell(s_t,a_t)\mathbb{V}_\ell(s_t,a_t)}\mathbf{1}_{\Omega^\ell_{s_t,a_t}}\right] \tag{8}$$

$$+ 272SB_\star\mathbb{E}\left[\sum_{t=1}^{T_M} A_\ell(s_t,a_t)\mathbf{1}_{\Omega^\ell_{s_t,a_t}}\right] \tag{9}$$

$$+ 2SB_\star\mathbb{E}\left[\sum_{t=1}^{T_M}\left(\mathbf{1}_{\{\theta_*(\cdot|s_t,a_t)\notin B_\ell(s_t,a_t)\}} + \mathbf{1}_{\{\theta_\ell(\cdot|s_t,a_t)\notin B_\ell(s_t,a_t)\}}\right)\right]. \tag{10}$$

The inner sum in (9) is bounded by $6SA\log^2(SAT_M/\delta)$ (see Lemma 9). To bound (10), we first show that $B_\ell(s,a)$ contains the true transition probability $\theta_*(\cdot|s,a)$ with high probability:

Lemma 7. *For any epoch ℓ and any state-action pair $(s,a) \in \mathcal{S}\times\mathcal{A}$, $\theta_*(\cdot|s,a) \in B_\ell(s,a)$ with probability at least $1 - \frac{\delta}{2SAn_\ell^+(s,a)}$.*

Proof. Fix $(s,a,s') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}^+$ and $0 < \delta' < 1$ (to be chosen later). Let $(Z_i)_{i=1}^\infty$ be a sequence of random variables drawn from the probability distribution $\theta_*(\cdot|s,a)$. Apply Lemma 8 below with $X_i = \mathbf{1}_{\{Z_i=s'\}}$ and $\delta_t = \frac{\delta'}{4St^2}$ to a prefix of length t of the sequence $(X_i)_{i=1}^\infty$, and apply union bound over all t and s' to obtain

$$\left|\hat\theta_\ell(s'|s,a) - \theta_*(s'|s,a)\right| \leq 2\sqrt{\frac{\hat\theta_\ell(s'|s,a)\log\frac{8Sn_\ell^{+2}(s,a)}{\delta'}}{n_\ell^+(s,a)}} + 7\log\frac{8Sn_\ell^{+2}(s,a)}{\delta'}$$

with probability at least $1-\delta'/2$ for all $s' \in \mathcal{S}^+$ and $\ell \geq 1$, simultaneously. Choose $\delta' = \delta/SAn_\ell^+(s,a)$ and use $S \geq 2$, $A \geq 2$ to complete the proof. □

Lemma 8 (Theorem D.3 (Anytime Bernstein) of (Rosenberg et al., 2020))**.** *Let $(X_n)_{n=1}^\infty$ be a sequence of independent and identically distributed random variables with expectation μ. Suppose that $0 \leq X_n \leq B$ almost surely. Then with probability at least $1-\delta$, the following holds for all $n \geq 1$ simultaneously:*

$$\left|\sum_{i=1}^n (X_i - \mu)\right| \leq 2\sqrt{B\sum_{i=1}^n X_i \log\frac{2n}{\delta}} + 7B\log\frac{2n}{\delta}.$$

Now, by rewriting the sum in (10) over epochs, we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T_M}\left(\mathbf{1}_{\{\theta_*(\cdot|s_t,a_t)\notin B_\ell(s_t,a_t)\}} + \mathbf{1}_{\{\theta_\ell(\cdot|s_t,a_t)\notin B_\ell(s_t,a_t)\}}\right)\right] \\
&= \mathbb{E}\left[\sum_{\ell=1}^{L_M}\sum_{t=t_\ell}^{t_{\ell+1}-1}\left(\mathbf{1}_{\{\theta_*(\cdot|s_t,a_t)\notin B_\ell(s_t,a_t)\}} + \mathbf{1}_{\{\theta_\ell(\cdot|s_t,a_t)\notin B_\ell(s_t,a_t)\}}\right)\right] \\
&= \sum_{s,a}\mathbb{E}\left[\sum_{\ell=1}^{L_M}\sum_{t=t_\ell}^{t_{\ell+1}-1}\mathbf{1}_{\{s_t=s,a_t=a\}}\left(\mathbf{1}_{\{\theta_*(\cdot|s,a)\notin B_\ell(s,a)\}} + \mathbf{1}_{\{\theta_\ell(\cdot|s,a)\notin B_\ell(s,a)\}}\right)\right] \\
&= \sum_{s,a}\mathbb{E}\left[\sum_{\ell=1}^{L_M}\left(n_{t_{\ell+1}}(s,a) - n_{t_\ell}(s,a)\right)\left(\mathbf{1}_{\{\theta_*(\cdot|s,a)\notin B_\ell(s,a)\}} + \mathbf{1}_{\{\theta_\ell(\cdot|s,a)\notin B_\ell(s,a)\}}\right)\right].
\end{aligned}$$

Note that $n_{t_{\ell+1}}(s,a) - n_{t_\ell}(s,a) \leq n_{t_\ell}(s,a)+1$ by the second stopping criterion. Moreover, observe that $B_\ell(s,a)$ is $\mathcal{F}_{t_\ell}$ measurable. Thus, it follows from the property of posterior sampling (Lemma 1) that $\mathbb{E}[\mathbf{1}_{\{\theta_\ell(\cdot|s,a)\notin B_\ell(s,a)\}}|\mathcal{F}_{t_\ell}] =$

$\mathbb{E}[\mathbf{1}_{\{\theta_*(\cdot|s,a)\notin B_\ell(s,a)\}}|\mathcal{F}_{t_\ell}] = \mathbb{P}(\theta_*(\cdot|s,a) \notin B_\ell(s,a)|\mathcal{F}_{t_\ell}) \le \delta/(2SAn_\ell^+(s,a))$, where the inequality is by Lemma 7. Using Monotone Convergence Theorem and that $\mathbf{1}_{\{m(t_\ell)\le M\}}$ is $\mathcal{F}_{t_\ell}$ measurable, we can write

$$\begin{aligned}
&\sum_{s,a}\mathbb{E}\left[\sum_{\ell=1}^{L_M}\left(n_{t_{\ell+1}}(s,a)-n_{t_\ell}(s,a)\right)\left(\mathbf{1}_{\{\theta_*(\cdot|s,a)\notin B_\ell(s,a)\}}+\mathbf{1}_{\{\theta_\ell(\cdot|s,a)\notin B_\ell(s,a)\}}\right)\right]\\
&\le \sum_{s,a}\sum_{\ell=1}^{\infty}\mathbb{E}\left[\mathbf{1}_{\{m(t_\ell)\le M\}}\left(n_{t_\ell}(s,a)+1\right)\mathbb{E}\left[\mathbf{1}_{\{\theta_*(\cdot|s,a)\notin B_\ell(s,a)\}}+\mathbf{1}_{\{\theta_\ell(\cdot|s,a)\notin B_\ell(s,a)\}}|\mathcal{F}_{t_\ell}\right]\right]\\
&\le \sum_{s,a}\sum_{\ell=1}^{\infty}\mathbb{E}\left[\mathbf{1}_{\{m(t_\ell)\le M\}}\left(n_{t_\ell}(s,a)+1\right)\frac{\delta}{SAn_\ell^+(s,a)}\right]\\
&\le 2\delta\mathbb{E}[L_M],
\end{aligned}$$

where the last inequality is by $n_{t_\ell}(s,a)+1 \le 2n_\ell^+(s,a)$ and Monotone Convergence Theorem.

We proceed by bounding (8). Denote by t_m the start time of interval m, define $t_{M+1} := T_M + 1$, and rewrite the sum in (8) over intervals to get

$$\mathbb{E}\left[\sum_{t=1}^{T_M}\sqrt{A_\ell(s_t,a_t)\mathbb{V}_\ell(s_t,a_t)}\mathbf{1}_{\Omega^\ell_{s_t,a_t}}\right] = \sum_{m=1}^{M}\mathbb{E}\left[\sum_{t=t_m}^{t_{m+1}-1}\sqrt{A_\ell(s_t,a_t)\mathbb{V}_\ell(s_t,a_t)}\mathbf{1}_{\Omega^\ell_{s_t,a_t}}\right]$$

Applying Cauchy-Schwartz twice on the inner expectation implies

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=t_m}^{t_{m+1}-1}\sqrt{A_\ell(s_t,a_t)\mathbb{V}_\ell(s_t,a_t)}\mathbf{1}_{\Omega^\ell_{s_t,a_t}}\right]\\
&\le \mathbb{E}\left[\sqrt{\sum_{t=t_m}^{t_{m+1}-1}A_\ell(s_t,a_t)}\cdot\sqrt{\sum_{t=t_m}^{t_{m+1}-1}\mathbb{V}_\ell(s_t,a_t)\mathbf{1}_{\Omega^\ell_{s_t,a_t}}}\right]\\
&\le \sqrt{\mathbb{E}\left[\sum_{t=t_m}^{t_{m+1}-1}A_\ell(s_t,a_t)\right]}\cdot\sqrt{\mathbb{E}\left[\sum_{t=t_m}^{t_{m+1}-1}\mathbb{V}_\ell(s_t,a_t)\mathbf{1}_{\Omega^\ell_{s_t,a_t}}\right]}\\
&\le 7B_\star\sqrt{\mathbb{E}\left[\sum_{t=t_m}^{t_{m+1}-1}A_\ell(s_t,a_t)\right]},
\end{aligned}$$

where the last inequality is by Lemma 10. Summing over M intervals and applying Cauchy-Schwartz, we get

$$\begin{aligned}
\sum_{m=1}^{M}\mathbb{E}\left[\sum_{t=t_m}^{t_{m+1}-1}\sqrt{A_\ell(s_t,a_t)\mathbb{V}_\ell(s_t,a_t)}\mathbf{1}_{\Omega^\ell_{s_t,a_t}}\right] &\le 7B_\star\sum_{m=1}^{M}\sqrt{\mathbb{E}\left[\sum_{t=t_m}^{t_{m+1}-1}A_\ell(s_t,a_t)\right]}\\
&\le 7B_\star\sqrt{M\sum_{m=1}^{M}\mathbb{E}\left[\sum_{t=t_m}^{t_{m+1}-1}A_\ell(s_t,a_t)\right]}\\
&= 7B_\star\sqrt{M\mathbb{E}\left[\sum_{t=1}^{T_M}A_\ell(s_t,a_t)\right]}\\
&\le 18B_\star\sqrt{MSA\mathbb{E}\left[\log^2\frac{SAT_M}{\delta}\right]},
\end{aligned}$$

where the last inequality follows from Lemma 9. Substituting these bounds in (8), (9), and (10), concavity of $\log^2 x$ for $x \ge 3$, and applying Jensen's inequality completes the proof.

Lemma 9. $\sum_{t=1}^{T_M} A_\ell(s_t, a_t) \leq 6SA\log^2(SAT_M/\delta)$.

Proof. Recall $A_\ell(s,a) = \frac{\log(SAn_\ell^+(s,a)/\delta)}{n_\ell^+(s,a)}$. Denote by $L := \log(SAT_M/\delta)$, an upper bound on the numerator of $A_\ell(s_t, a_t)$. we have

$$\begin{aligned}
\sum_{t=1}^{T_M} A_\ell(s_t,a_t) &\leq \sum_{t=1}^{T_M} \frac{L}{n_\ell^+(s_t,a_t)} = L\sum_{s,a}\sum_{t=1}^{T_M}\frac{\mathbf{1}_{\{s_t=s,a_t=a\}}}{n_\ell^+(s,a)} \\
&\leq 2L\sum_{s,a}\sum_{t=1}^{T_M}\frac{\mathbf{1}_{\{s_t=s,a_t=a\}}}{n_t^+(s,a)} = 2L\sum_{s,a}\mathbf{1}_{\{n_{T_M+1}(s,a)>0\}} + 2L\sum_{s,a}\sum_{j=1}^{n_{T_M+1}(s,a)-1}\frac{1}{j} \\
&\leq 2LSA + 2L\sum_{s,a}(1+\log n_{T_M+1}(s,a)) \\
&\leq 4LSA + 2LSA\log T_M \leq 6LSA\log T_M.
\end{aligned}$$

Here the second inequality is by $n_\ell^+(s,a) \geq 0.5n_t^+(s,a)$ (the second criterion in determining the epoch length), the third inequality is by $\sum_{x=1}^n 1/x \leq 1 + \log n$, and the fourth inequality is by $n_{T_M+1}(s,a) \leq T_M$. The proof is complete by noting that $\log T_M \leq L$. □

Lemma 10. *For any interval m, $\mathbb{E}[\sum_{t=t_m}^{t_{m+1}-1}\mathbb{V}_\ell(s_t,a_t)\mathbf{1}_{\Omega^\ell}] \leq 44B_\star^2$.*

Proof. To proceed with the proof, we need the following two technical lemmas.

Lemma 11. *Let (s,a) be a known state-action pair and m be an interval. If $\Omega_{s,a}^\ell$ holds, then for any state $s' \in \mathcal{S}^+$,*

$$|\theta_*(s'|s,a) - \theta_\ell(s'|s,a)| \leq \frac{1}{8}\sqrt{\frac{c_{\min}\theta_*(s'|s,a)}{SB_\star}} + \frac{c_{\min}}{4SB_\star}.$$

Proof. From Lemma 6, we know that if $\Omega_{s,a}^\ell$ holds, then

$$|\theta_*(s'|s,a) - \theta_\ell(s'|s,a)| \leq 8\sqrt{\theta_*(s'|s,a)A_\ell(s,a)} + 136A_\ell(s,a),$$

with $A_\ell(s,a) = \frac{\log(SAn_\ell^+(s,a)/\delta)}{n_\ell^+(s,a)}$. The proof is complete by noting that $\log(x)/x$ is decreasing, and that $n_\ell^+(s,a) \geq \alpha\cdot\frac{B_\star S}{c_{\min}}\log\frac{B_\star SA}{\delta c_{\min}}$ for some large enough constant α since (s,a) is known. □

Lemma 12 (Lemma B.15. of Rosenberg et al. (2020))**.** *Let $(X_t)_{t=1}^\infty$ be a martingale difference sequence adapted to the filtration $(\mathcal{F}_t)_{t=0}^\infty$. Let $Y_n = (\sum_{t=1}^n X_t)^2 - \sum_{t=1}^n \mathbb{E}[X_t^2|\mathcal{F}_{t-1}]$. Then $(Y_n)_{n=0}^\infty$ is a martingale, and in particular if τ is a stopping time such that $\tau \leq c$ almost surely, then $\mathbb{E}[Y_\tau] = 0$.*

By the definition of the intervals, all the state-action pairs within an interval except possibly the first one are known. Therefore, we bound

$$\mathbb{E}\left[\sum_{t=t_m}^{t_{m+1}-1}\mathbb{V}_\ell(s_t,a_t)\mathbf{1}_{\Omega_{s_t,a_t}^\ell}\Big|\mathcal{F}_{t_m}\right] = \mathbb{E}\left[\mathbb{V}_\ell(s_{t_m},a_{t_m})\mathbf{1}_{\Omega_{s_t,a_t}^\ell}|\mathcal{F}_{t_m}\right] + \mathbb{E}\left[\sum_{t=t_m+1}^{t_{m+1}-1}\mathbb{V}_\ell(s_t,a_t)\mathbf{1}_{\Omega_{s_t,a_t}^\ell}\Big|\mathcal{F}_{t_m}\right].$$

The first summand is upper bounded by $B_\star^2$. To bound the second term, define $Z_\ell^t := [V(s_t';\theta_\ell) - \sum_{s'\in\mathcal{S}}\theta_*(s'|s_t,a_t)V(s';\theta_\ell)]\mathbf{1}_{\Omega_{s_t,a_t}^\ell}$. Conditioned on $\mathcal{F}_{t_m}, \theta_*$ and θ_ℓ, $(Z_\ell^t)_{t\geq t_m}$ constitutes a martingale difference sequence with respect to the filtration $(\mathcal{F}_{t+1}^m)_{t\geq t_m}$, where $\mathcal{F}_t^m$ is the sigma algebra generated by $\{(s_{t_m},a_{t_m}),\cdots,(s_t,a_t)\}$.

Moreover, $t_{m+1}-1$ is a stopping time with respect to $(\mathcal{F}^m_{t+1})_{t\geq t_m}$ and is bounded by $t_m + 2B_\star/c_{\min}$. Therefore, Lemma 12 implies that

$$\mathbb{E}\left[\sum_{t=t_m+1}^{t_{m+1}-1} \mathbb{V}_\ell(s_t,a_t)\mathbf{1}_{\Omega^\ell_{s_t,a_t}}\middle|\mathcal{F}_{t_m},\theta_*,\theta_\ell\right] = \mathbb{E}\left[\left(\sum_{t=t_m+1}^{t_{m+1}-1} Z_\ell^t \mathbf{1}_{\Omega^\ell_{s_t,a_t}}\right)^2\middle|\mathcal{F}_{t_m},\theta_*,\theta_\ell\right]. \tag{11}$$

We proceed by bounding $|\sum_{t=t_m+1}^{t_{m+1}-1} Z_\ell^t\mathbf{1}_{\Omega^\ell_{s_t,a_t}}|$ in terms of $\sum_{t=t_m+1}^{t_{m+1}-1}\mathbb{V}_\ell(s_t,a_t)\mathbf{1}_{\Omega^\ell_{s_t,a_t}}$ and combine with the left hand side to complete the proof. We have

$$\begin{aligned}
\left|\sum_{t=t_m+1}^{t_{m+1}-1} Z_\ell^t\mathbf{1}_{\Omega^\ell_{s_t,a_t}}\right| &= \left|\sum_{t=t_m+1}^{t_{m+1}-1}\left[V(s_t';\theta_\ell) - \sum_{s'\in\mathcal{S}}\theta_*(s'|s_t,a_t)V(s';\theta_\ell)\right]\mathbf{1}_{\Omega^\ell_{s_t,a_t}}\right| \\
&\leq \left|\sum_{t=t_m+1}^{t_{m+1}-1}[V(s_t';\theta_\ell) - V(s_t;\theta_\ell)]\right| \qquad (12)\\
&+ \left|\sum_{t=t_m+1}^{t_{m+1}-1}\left[V(s_t;\theta_\ell) - \sum_{s'\in\mathcal{S}}\theta_\ell(s'|s_t,a_t)V(s';\theta_\ell)\right]\right| \qquad (13)\\
&+ \left|\sum_{t=t_m+1}^{t_{m+1}-1}\sum_{s'\in\mathcal{S}}[\theta_\ell(s'|s_t,a_t) - \theta_*(s'|s_t,a_t)]\left(V(s';\theta_\ell) - \sum_{s''\in\mathcal{S}^+}\theta_*(s''|s_t,a_t)V(s'';\theta_\ell)\right)\mathbf{1}_{\Omega^\ell_{s_t,a_t}}\right|. \qquad (14)
\end{aligned}$$

where (14) is by the fact that $\theta_\ell(\cdot|s_t,a_t),\theta_*(\cdot|s_t,a_t)$ are probability distributions and $\sum_{s''\in\mathcal{S}^+}\theta_*(s''|s_t,a_t)V(s'';\theta_\ell)$ is independent of s' and $V(g;\theta_\ell)=0$. (12) is a telescopic sum (recall that $s_{t+1}=s_t'$ if $s_t'\neq g$) and is bounded by $B_\star$. It follows from the Bellman equation that (13) is equal to $\sum_{t=t_m+1}^{t_{m+1}-1} c(s_t,a_t)$. By definition, the interval ends as soon as the cost accumulates to $B_\star$ during the interval. Moreover, since $V(\cdot;\theta_\ell)\leq B_\star$, the algorithm does not choose an action with instantaneous cost more than $B_\star$. This implies that $\sum_{t=t_m+1}^{t_{m+1}-1} c(s_t,a_t)\leq 2B_\star$. To bound (14) we use the Bernstein confidence set, but taking into account that all the state-action pairs in the summation are known, we can use Lemma 11 to obtain

$$\begin{aligned}
&\sum_{s'\in\mathcal{S}}(\theta_\ell(s'|s_t,a_t) - \theta_*(s'|s_t,a_t))\left(V(s';\theta_\ell) - \sum_{s''\in\mathcal{S}^+}\theta_*(s''|s_t,a_t)V(s'';\theta_\ell)\right)\mathbf{1}_{\Omega^\ell_{s_t,a_t}} \\
&\leq \sum_{s'\in\mathcal{S}}\frac{1}{8}\sqrt{\frac{c_{\min}\theta_*(s'|s_t,a_t)\left(V(s';\theta_\ell) - \sum_{s''\in\mathcal{S}^+}\theta_*(s''|s_t,a_t)V(s'';\theta_\ell)\right)^2\mathbf{1}_{\Omega^\ell_{s_t,a_t}}}{SB_\star}} \\
&\quad + \sum_{s'\in\mathcal{S}}\frac{c_{\min}}{4SB_\star}\left|V(s';\theta_\ell) - \sum_{s''\in\mathcal{S}^+}\theta_*(s''|s_t,a_t)V(s'';\theta_\ell)\right| \\
&\leq \frac{1}{4}\sqrt{\frac{c_{\min}\mathbb{V}_\ell(s_t,a_t)\mathbf{1}_{\Omega^\ell_{s_t,a_t}}}{B_\star}} + \frac{c(s_t,a_t)}{2}.
\end{aligned}$$

The last inequality follows from Cauchy-Schwartz inequality, $|\mathcal{S}^+|\leq 2S$, $|V(\cdot;\theta_\ell)|\leq B_\star$, and $c_{\min}\leq c(s_t,a_t)$. Summing over the time steps in interval m and applying Cauchy-Schwartz, we get

$$\begin{aligned}
\sum_{t=t_m+1}^{t_{m+1}-1}\left[\frac{1}{4}\sqrt{\frac{c_{\min}\mathbb{V}_\ell(s_t,a_t)\mathbf{1}_{\Omega^\ell_{s_t,a_t}}}{B_\star}} + \frac{c(s_t,a_t)}{2}\right] &\leq \frac{1}{4}\sqrt{(t_{m+1}-t_m)\frac{c_{\min}\sum_{t=t_m+1}^{t_{m+1}-1}\mathbb{V}_\ell(s_t,a_t)\mathbf{1}_{\Omega^\ell_{s_t,a_t}}}{B_\star}} \\
&\quad + \frac{\sum_{t=t_m+1}^{t_{m+1}-1}c(s_t,a_t)}{2} \\
&\leq \frac{1}{4}\sqrt{2\sum_{t=t_m+1}^{t_{m+1}-1}\mathbb{V}_\ell(s_t,a_t)\mathbf{1}_{\Omega^\ell_{s_t,a_t}}} + B_\star.
\end{aligned}$$

The last inequality follows from the fact that duration of interval m is at most $2B_\star/c_{\min}$ and its cumulative cost is at most $2B_\star$. Substituting these bounds into (11) implies that

$$\mathbb{E}\left[\sum_{t=t_m+1}^{t_{m+1}-1} \mathbb{V}_\ell(s_t,a_t)\mathbf{1}_{\Omega^\ell_{s_t,a_t}}\Big|\mathcal{F}_{t_m},\theta_*,\theta_\ell\right] \le \mathbb{E}\left[\left(4B_\star + \frac{1}{4}\sqrt{2\sum_{t=t_m+1}^{t_{m+1}-1} \mathbb{V}_\ell(s_t,a_t)\mathbf{1}_{\Omega^\ell_{s_t,a_t}}}\right)^2\Big|\mathcal{F}_{t_m},\theta_*,\theta_\ell\right]$$
$$\le 32B_\star^2 + \frac{1}{4}\mathbb{E}\left[\sum_{t=t_m+1}^{t_{m+1}-1} \mathbb{V}_\ell(s_t,a_t)\mathbf{1}_{\Omega^\ell_{s_t,a_t}}\Big|\mathcal{F}_{t_m},\theta_*,\theta_\ell\right],$$

where the last inequality is by $(a+b)^2 \le 2(a^2+b^2)$ with $b = \frac{1}{4}\sqrt{2\sum_{t=t_m+1}^{t_{m+1}-1}\mathbb{V}_\ell(s_t,a_t)\mathbf{1}_{\Omega^\ell_{s_t,a_t}}}$ and $a = 4B_\star$. Rearranging implies that $\mathbb{E}\left[\sum_{t=t_m+1}^{t_{m+1}-1}\mathbb{V}_\ell(s_t,a_t)\mathbf{1}_{\Omega^\ell_{s_t,a_t}}|\mathcal{F}_{t_m},\theta_*,\theta_\ell\right] \le 43B_\star^2$ and the proof is complete. □

□

B.5. Proof of Theorem 1

Theorem (restatement of Theorem 1)**.** Suppose Assumptions 1 and 2 hold. Then, the regret bound of the `PSRL-SSP` algorithm is bounded as

$$R_K = \mathcal{O}\left(B_\star S\sqrt{KA}L^2 + S^2A\sqrt{\frac{{B_\star}^3}{c_{\min}}}L^2\right),$$

where $L = \log(B_\star SAKc_{\min}^{-1})$.

Proof. Denote by C_M the total cost after M intervals. Recall that

$$\mathbb{E}[C_M] = K\mathbb{E}[V(s_{\text{init}};\theta_*)] + R_M = K\mathbb{E}[V(s_{\text{init}};\theta_*)] + R_M^1 + R_M^2 + R_M^3$$

Using Lemmas 3, 4, and 5 with $\delta = 1/K$ obtains

$$\mathbb{E}[C_M] \le K\mathbb{E}[V(s_{\text{init}};\theta_*)] + \mathcal{O}\left(B_\star\mathbb{E}[L_M] + B_\star S\sqrt{MA\log^2(SAK\mathbb{E}[T_M])} + B_\star S^2A\log^2(SAK\mathbb{E}[T_M])\right). \tag{15}$$

Recall that $L_M \le \sqrt{2SAK\log T_M} + SA\log T_M$. Taking expectation from both sides and using Jensen's inequality gets us $\mathbb{E}[L_M] \le \sqrt{2SAK\log\mathbb{E}[T_M]} + SA\log\mathbb{E}[T_M]$. Moreover, taking expectation from both sides of (3), plugging in the bound on $\mathbb{E}[L_M]$, and concavity of $\log(x)$ implies

$$M \le \frac{\mathbb{E}[C_M]}{B_\star} + K + \sqrt{2SAK\log\mathbb{E}[T_M]} + SA\log\mathbb{E}[T_M] + \mathcal{O}\left(\frac{B_\star S^2A}{c_{\min}}\log\frac{B_\star KSA}{c_{\min}}\right).$$

Substituting this bound in (15), using subadditivity of the square root, and simplifying yields

$$\mathbb{E}[C_M] \le K\mathbb{E}[V(s_{\text{init}};\theta_*)] + \mathcal{O}\Bigg(B_\star S\sqrt{KA\log^2(SAK\mathbb{E}[T_M])} + S\sqrt{B_\star\mathbb{E}[C_M]A\log^2(SAK\mathbb{E}[T_M])}$$
$$+ B_\star S^{\frac{5}{4}}A^{\frac{3}{4}}K^{\frac{1}{4}}\log^{\frac{5}{4}}(SAK\mathbb{E}[T_M]) + S^2A\sqrt{\frac{B_\star^3}{c_{\min}}\log^3\frac{B_\star SAK\mathbb{E}[T_M]}{c_{\min}}}\Bigg).$$

Solving for $\mathbb{E}[C_M]$ (by using the primary inequality that $x \leq a\sqrt{x}+b$ implies $x \leq (a+\sqrt{b})^2$ for $a, b > 0$), using $K \geq S^2 A$, $V(s_{\text{init}}; \theta_*) \leq B_\star$, and simplifying the result gives

$$
\begin{aligned}
\mathbb{E}[C_M] \leq & \Bigg(\mathcal{O}\left(S\sqrt{B_\star A \log^2(SAK\mathbb{E}[T_M])} \right) \\
& + \sqrt{K\mathbb{E}[V(s_{\text{init}};\theta_*)] + \mathcal{O}\left(B_\star S\sqrt{KA\log^{2.5}(SAK\mathbb{E}[T_M])} + S^2 A\sqrt{\frac{B_\star^3}{c_{\min}} \log^3 \frac{B_\star SAK\mathbb{E}[T_M]}{c_{\min}}} \right)} \Bigg)^2 \\
\leq & \; \mathcal{O}\left(B_\star S^2 A \log^2 \frac{SA\mathbb{E}[T_M]}{\delta} \right) \\
& + K\mathbb{E}[V(s_{\text{init}};\theta_*)] + \mathcal{O}\Bigg(B_\star S\sqrt{KA\log^{2.5}(SAK\mathbb{E}[T_M])} + S^2 A\sqrt{\frac{B_\star^3}{c_{\min}} \log^3 \frac{B_\star SAK\mathbb{E}[T_M]}{c_{\min}}} \\
& + B_\star S\sqrt{KA\log^4(SAK\mathbb{E}[T_M])} + S^2 A\left(\frac{{B_\star}^5}{c_{\min}} \log^7 \frac{B_\star SAK\mathbb{E}[T_M]}{c_{\min}} \right)^{\frac{1}{4}} \Bigg) \\
\leq & \; K\mathbb{E}[V(s_{\text{init}};\theta_*)] + \mathcal{O}\left(B_\star S\sqrt{KA\log^4 SAK\mathbb{E}[T_M])} + S^2 A\sqrt{\frac{B_\star^3}{c_{\min}} \log^4 \frac{B_\star SAK\mathbb{E}[T_M]}{c_{\min}}} \right). \qquad (16)
\end{aligned}
$$

Note that by simplifying this bound, we can write $\mathbb{E}[C_M] \leq \mathcal{O}\left(\sqrt{{B_\star}^3 S^4 A^2 K^2 \mathbb{E}[T_M]/c_{\min}} \right)$. On the other hand, we have that $c_{\min} T_M \leq C_M$ which implies $\mathbb{E}[T_M] \leq \mathbb{E}[C_M]/c_{\min}$. Isolating $\mathbb{E}[T_M]$ implies $\mathbb{E}[T_M] \leq \mathcal{O}\left({B_\star}^3 S^4 A^2 K^2 / c_{\min}^3 \right)$. Substituting this bound into (16) yields

$$
\mathbb{E}[C_M] \leq K\mathbb{E}[V(s_{\text{init}};\theta_*)] + \mathcal{O}\left(B_\star S\sqrt{KA\log^4 \frac{B_\star SAK}{c_{\min}}} + S^2 A\sqrt{\frac{B_\star^3}{c_{\min}} \log^4 \frac{B_\star SAK}{c_{\min}}} \right).
$$

We note that this bound holds for any number of M intervals as long as the K episodes have not elapsed. Since, $c_{\min} > 0$, this implies that the K episodes eventually terminate and the claimed bound of the theorem for R_K holds. □

B.6. Proof of Theorem 2

Theorem (restatement of Theorem 2)**.** Suppose Assumption 1 holds. Running the `PSRL-SSP` algorithm with costs $c_\epsilon(s,a) := \max\{c(s,a), \epsilon\}$ for $\epsilon = (S^2 A/K)^{2/3}$ yields

$$
R_K = \mathcal{O}\left(B_\star S\sqrt{KA}\tilde{L}^2 + (S^2 A)^{\frac{2}{3}} K^{\frac{1}{3}} (B_\star^{\frac{3}{2}} \tilde{L}^2 + T_\star) + S^2 A T_\star^{\frac{3}{2}} \tilde{L}^2 \right),
$$

where $\tilde{L} := \log(KB_\star T_\star SA)$ and $T_\star$ is an upper bound on the expected time the optimal policy takes to reach the goal from any initial state.

Proof. Denote by T_K^ϵ the time to complete K episodes if the algorithm runs with the perturbed costs $c_\epsilon(s,a)$ and let $V_\epsilon(s_{\text{init}};\theta_*)$, $V_\epsilon^\pi(s_{\text{init}};\theta_*)$ be the optimal value function and the value function for policy π in the SSP with cost function $c_\epsilon(s,a)$ and transition kernel θ_*. We can write

$$
\begin{aligned}
R_K &= \mathbb{E}\left[\sum_{t=1}^{T_K^\epsilon} c(s_t, a_t) - KV(s_{\text{init}};\theta_*) \right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T_K^\epsilon} c_\epsilon(s_t, a_t) - KV(s_{\text{init}};\theta_*) \right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T_K^\epsilon} c_\epsilon(s_t, a_t) - KV_\epsilon(s_{\text{init}};\theta_*) \right] + K\mathbb{E}\left[V_\epsilon(s_{\text{init}};\theta_*) - V(s_{\text{init}};\theta_*) \right]. \qquad (17)
\end{aligned}
$$

Theorem 1 implies that the first term is bounded by

$$\mathbb{E}\left[\sum_{t=1}^{T_K^\epsilon} c_\epsilon(s_t, a_t) - KV_\epsilon(s_{\text{init}}; \theta_*)\right] = \mathcal{O}\left(B_\star^\epsilon S\sqrt{KA}L_\epsilon^2 + S^2A\sqrt{\frac{{B_\star^\epsilon}^3}{\epsilon}}L_\epsilon^2\right),$$

with $L_\epsilon = \log(B_\star^\epsilon SAK/\epsilon)$ and $B_\star^\epsilon \leq B_\star + \epsilon T_\star$ (to see this note that $V_\epsilon(s;\theta_*) \leq V_\epsilon^{\pi^*}(s;\theta_*) \leq B_\star + \epsilon T_\star$). To bound the second term of (17), we have

$$V_\epsilon(s_{\text{init}}; \theta_*) \leq V_\epsilon^{\pi^*}(s_{\text{init}}; \theta_*) \leq V(s_{\text{init}}; \theta_*) + \epsilon T_\star.$$

Combining these bounds, we can write

$$R_K = \mathcal{O}\left(B_\star S\sqrt{KA}L_\epsilon^2 + \epsilon T_\star S\sqrt{KA}L_\epsilon^2 + S^2A\sqrt{\frac{(B_\star + \epsilon T_\star)^3}{\epsilon}}L_\epsilon^2 + KT_\star\epsilon\right).$$

Substituting $\epsilon = (S^2A/K)^{2/3}$, and simplifying the result with $K \geq S^2A$ and $B_\star \leq T_\star$ (since $c(s,a) \leq 1$) implies

$$R_K = \mathcal{O}\left(B_\star S\sqrt{KA}\tilde{L}^2 + (S^2A)^{\frac{2}{3}}K^{\frac{1}{3}}(B_\star^{\frac{3}{2}}\tilde{L}^2 + T_\star) + S^2AT_\star^{\frac{3}{2}}\tilde{L}^2\right),$$

where $\tilde{L} = \log(KB_\star T_\star SA)$. This completes the proof. $\square$