Posterior Sampling-Based Online Learning for the Stochastic Shortest Path Model (Supplementary Material)

Mehdi Jafarnia-Jahromi[1] **Liyu Chen**[3] **Rahul Jain**[2,3,4] **Haipeng Luo**[3]

[1]Google DeepMind
[2]ECE Department, University of Southern California
[3]CS Department, University of Southern California
[4]USC Center for Autonomy and AI

A PROOFS

A.1 PROOF OF LEMMA 4.2

Lemma (restatement of Lemma 4.2)**.** The number of epochs is bounded as $L_M \leq \sqrt{2SAK \log T_M} + SA \log T_M$.

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,

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

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

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

where the second inequality follows from Cauchy-Schwarz. 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

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

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

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

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

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

A.2 PROOF OF LEMMA 4.3

Lemma (restatement of Lemma 4.3)**.** 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 2.1. □

A.3 PROOF OF LEMMA 4.4

Lemma (restatement of Lemma 4.4)**.** 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

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

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 $K_\ell \leq K_{\ell-1} + 1$ for all ℓ. Thus, each term in the summation can be bounded as

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

$\mathbf{1}_{\{m(t_\ell) \leq M\}}(K_{\ell-1}+1)$ is $\mathcal{F}_{t_\ell}$ measurable. Therefore, applying the property of posterior sampling (Lemma 4.1) implies

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

Substituting this into R_M^2, we obtain

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

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

A.4 PROOF OF LEMMA 4.5

Lemma (restatement of Lemma 4.5)**.** The third term R_M^3 can be bounded as

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

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

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

The last equality follows from Dominated Convergence Theorem, tower property of conditional expectation, and that $\mathbf{1}_{\{m(t)\leq M\}}$ 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

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

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 (4) 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 A.1. *Denote* $A_\ell(s,a) = \frac{\log(SAn_\ell^+(s,a)/\delta)}{n_\ell^+(s,a)}$. *If* $\Omega_{s,a}^\ell$ *holds, then*

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

Proof. Since $\Omega_{s,a}^\ell$ holds, by (4) we have that

$$
\widehat{\theta}_\ell(s'|s,a) - \theta_*(s'|s,a) \leq 4\sqrt{\widehat{\theta}_\ell(s'|s,a)A_\ell(s,a)} + 28A_\ell(s,a).
$$

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

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

where the last inequality is by sub-linearity of the square root. Substituting this bound into (4) yields

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

Similarly,

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

Using the triangle inequality completes the proof. □

Note that if either of $\theta_*(\cdot|s_t,a_t)$ or $\theta_\ell(\cdot|s_t,a_t)$ is not in $B_\ell(s_t,a_t)$, then the inner term of (1) can be bounded by $2SB_\star$ (note that $|\mathcal{S}^+| \le 2S$ and $V(\cdot;\theta_\ell) \le B_\star$). Thus, applying Lemma A.1 implies that

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

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 (5). Here the last inequality follows from Cauchy-Schwarz, $|\mathcal{S}^+| \le 2S$, $V(\cdot;\theta_\ell) \le B_\star$ and the definition of $\mathbb{V}_\ell$. Substituting this into (1) yields

$$R^3_M \le 16\sqrt{S}\mathbb{E}\left[\sum_{t=1}^{T_M} \sqrt{A_\ell(s_t,a_t)\mathbb{V}_\ell(s_t,a_t)}\mathbf{1}_{\Omega^\ell_{s_t,a_t}} \right] \tag{2}$$

$$+ 272SB_\star\mathbb{E}\left[\sum_{t=1}^{T_M} A_\ell(s_t,a_t)\mathbf{1}_{\Omega^\ell_{s_t,a_t}} \right] \tag{3}$$

$$+ 2SB_\star\mathbb{E}\left[\sum_{t=1}^{T_M} \left(\mathbf{1}_{\{\theta_*(\cdot|s_t,a_t)\notin B_\ell(s_t,a_t)\}} + \mathbf{1}_{\{\theta_\ell(\cdot|s_t,a_t)\notin B_\ell(s_t,a_t)\}} \right) \right]. \tag{4}$$

The inner sum in (3) is bounded by $6SA\log^2(SAT_M/\delta)$ (see Lemma A.4). To bound (4), we first show that $B_\ell(s,a)$ contains the true transition probability $\theta_*(\cdot|s,a)$ with high probability:

Lemma A.2. *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 A.3 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

$$\left| \hat{\theta}_\ell(s'|s,a) - \theta_*(s'|s,a) \right| \le 2\sqrt{\frac{\hat{\theta}_\ell(s'|s,a)\log\frac{8Sn^{+2}_\ell(s,a)}{\delta'}}{n^+_\ell(s,a)}} + 7\log\frac{8Sn^{+2}_\ell(s,a)}{\delta'}$$

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 A.3 (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 (4) 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 4.1) that $\mathbb{E}[\mathbf{1}_{\{\theta_\ell(\cdot|s,a)\notin B_\ell(s,a)\}}|\mathcal{F}_{t_\ell}] = \mathbb{E}[\mathbf{1}_{\{\theta_*(\cdot|s,a)\notin B_\ell(s,a)\}}|\mathcal{F}_{t_\ell}] = \mathbb{P}(\theta_*(\cdot|s,a) \notin B_\ell(s,a)|\mathcal{F}_{t_\ell}) \leq \delta/(2SAn_\ell^+(s,a))$, where the inequality is by Lemma A.2. Using Monotone Convergence Theorem and that $\mathbf{1}_{\{m(t_\ell)\leq M\}}$ is $\mathcal{F}_{t_\ell}$ measurable, we can write

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

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

We proceed by bounding (2). Denote by t_m the start time of interval m, define $t_{M+1} := T_M + 1$, and rewrite the sum in (2) 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-Schwarz 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 A.5. Summing over M intervals and applying Cauchy-Schwarz, 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 A.4. Substituting these bounds in (2), (3), (4), concavity of $\log^2 x$ for $x \ge 3$, and applying Jensen's inequality completes the proof.

Lemma A.4. $\sum_{t=1}^{T_M} A_\ell(s_t,a_t) \le 6SA\log^2(SAT_M/\delta)$.

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

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

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

Lemma A.5. *For any interval m,* $\mathbb{E}[\sum_{t=t_m}^{t_{m+1}-1}\mathbb{V}_\ell(s_t,a_t)\mathbf{1}_{\Omega^\ell}] \le 44B_\star^2$.

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

Lemma A.6. *Let (s,a) be a known state-action pair and m be an interval. If $\Omega^\ell_{s,a}$ holds, then for any state $s' \in \mathcal{S}^+$,*

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

Proof. From Lemma A.1, we know that if $\Omega^\ell_{s,a}$ holds, then

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

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 $n^+_\ell(s,a) \ge \alpha \cdot \frac{B_\star S}{c_{\min}} \log \frac{B_\star SA}{\delta c_{\min}}$ for some large enough constant α since (s,a) is known. □

Lemma A.7 (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 $\tau \le c$ 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

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

The first summand is upper bounded by $B_\star^2$. To bound the second term, define $Z^t_\ell := [V(s'_t;\theta_\ell) - \sum_{s'\in\mathcal{S}} \theta_*(s'|s_t,a_t)V(s';\theta_\ell)]\mathbf{1}_{\Omega^\ell_{s_t,a_t}}$. Conditioned on $\mathcal{F}_{t_m}, \theta_*$ and θ_ℓ, $(Z^t_\ell)_{t\ge t_m}$ constitutes a martingale difference sequence with respect to the filtration $(\mathcal{F}^m_{t+1})_{t\ge t_m}$, where $\mathcal{F}^m_t$ 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 $(\mathcal{F}^m_{t+1})_{t\ge t_m}$ and is bounded by $t_m + 2B_\star/c_{\min}$. Therefore, Lemma A.7 implies that

$$\mathbb{E}\left[\sum_{t=t_m+1}^{t_{m+1}-1} \mathbb{V}_\ell(s_t,a_t)\mathbf{1}_{\Omega^\ell_{s_t,a_t}} \middle| \mathcal{F}_{t_m},\theta_*,\theta_\ell\right] = \mathbb{E}\left[\left(\sum_{t=t_m+1}^{t_{m+1}-1} Z^t_\ell \mathbf{1}_{\Omega^\ell_{s_t,a_t}}\right)^2 \middle| \mathcal{F}_{t_m},\theta_*,\theta_\ell\right]. \tag{5}$$

We proceed by bounding $|\sum_{t=t_m+1}^{t_{m+1}-1} Z^t_\ell \mathbf{1}_{\Omega^\ell_{s_t,a_t}}|$ 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

$$\left|\sum_{t=t_m+1}^{t_{m+1}-1} Z^t_\ell \mathbf{1}_{\Omega^\ell_{s_t,a_t}}\right| = \left|\sum_{t=t_m+1}^{t_{m+1}-1}\left[V(s'_t;\theta_\ell) - \sum_{s'\in\mathcal{S}}\theta_*(s'|s_t,a_t)V(s';\theta_\ell)\right]\mathbf{1}_{\Omega^\ell_{s_t,a_t}}\right|$$

$$\le \left|\sum_{t=t_m+1}^{t_{m+1}-1}[V(s'_t;\theta_\ell) - V(s_t;\theta_\ell)]\right| \tag{6}$$

$$+ \left|\sum_{t=t_m+1}^{t_{m+1}-1}\left[V(s_t;\theta_\ell) - \sum_{s'\in\mathcal{S}}\theta_\ell(s'|s_t,a_t)V(s';\theta_\ell)\right]\right| \tag{7}$$

$$+ \left|\sum_{t=t_m+1}^{t_{m+1}-1}\sum_{s'\in\mathcal{S}^+}[\theta_\ell(s'|s_t,a_t) - \theta_*(s'|s_t,a_t)]\left(V(s';\theta_\ell) - \sum_{s''\in\mathcal{S}^+}\theta_*(s''|s_t,a_t)V(s'';\theta_\ell)\right)\mathbf{1}_{\Omega^\ell_{s_t,a_t}}\right|. \tag{8}$$

where (8) 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 $V(g;\theta_\ell) = 0$. (6) 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 (7) 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 $V(\cdot;\theta_\ell) \le B_\star$, 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 (8) 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 A.6 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-Schwarz 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-Schwarz, 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 (5) implies that

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

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

□

A.5 PROOF OF THEOREM 3.5

Theorem (restatement of Theorem 3.5)**.** Suppose Assumptions 2.1 and 3.4 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 4.3, 4.4, and 4.5 with $\delta = 1/K$ obtains

$$\begin{aligned}\mathbb{E}[C_M] &\leq K\mathbb{E}[V(s_{\text{init}};\theta_*)] \\ &\quad + \mathcal{O}\left(B_\star \mathbb{E}[L_M] + B_\star S\sqrt{MA\log^2(SAK\mathbb{E}[T_M])} + B_\star S^2 A \log^2(SAK\mathbb{E}[T_M])\right). \end{aligned} \tag{9}$$

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

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

Substituting this bound in (9), using subadditivity of the square root, and simplifying yields

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

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

$$\begin{aligned}\mathbb{E}[C_M] &\leq \Bigg(\mathcal{O}\left(S\sqrt{B_\star A \log^2(SAK\mathbb{E}[T_M])}\right) \\ &\quad + \sqrt{K\mathbb{E}[V(s_{\text{init}};\theta_*)] + \mathcal{O}\left(B_\star S\sqrt{KA\log^{2.5}(SAK\mathbb{E}[T_M])} + S^2 A\sqrt{\frac{B_\star^3}{c_{\min}} \log^3 \frac{B_\star SAK\mathbb{E}[T_M]}{c_{\min}}}\right)}\Bigg)^2 \\ &\leq \mathcal{O}\left(B_\star S^2 A \log^2 \frac{SA\mathbb{E}[T_M]}{\delta}\right) \\ &\quad + K\mathbb{E}[V(s_{\text{init}};\theta_*)] + \mathcal{O}\Bigg(B_\star S\sqrt{KA\log^{2.5}(SAK\mathbb{E}[T_M])} + S^2 A\sqrt{\frac{B_\star^3}{c_{\min}} \log^3 \frac{B_\star SAK\mathbb{E}[T_M]}{c_{\min}}} \\ &\quad + B_\star S\sqrt{KA\log^4(SAK\mathbb{E}[T_M])} + S^2 A\left(\frac{{B_\star}^5}{c_{\min}} \log^7 \frac{B_\star SAK\mathbb{E}[T_M]}{c_{\min}}\right)^{\frac{1}{4}}\Bigg) \\ &\leq K\mathbb{E}[V(s_{\text{init}};\theta_*)] + \mathcal{O}\left(B_\star S\sqrt{KA\log^4 SAK\mathbb{E}[T_M])} + S^2 A\sqrt{\frac{B_\star^3}{c_{\min}} \log^4 \frac{B_\star SAK\mathbb{E}[T_M]}{c_{\min}}}\right). \end{aligned} \tag{10}$$

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 (10) 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. ☐

A.6 PROOF OF THEOREM 3.6

Theorem (restatement of Theorem 3.6)**.** Suppose Assumption 2.1 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 $\tilde{L} := \log(KB_\star T_\star SA)$ 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

$$\begin{aligned} R_K &= \mathbb{E}\left[\sum_{t=1}^{T_K^\epsilon} c(s_t,a_t) - KV(s_{\text{init}};\theta_*)\right] \\ &\le \mathbb{E}\left[\sum_{t=1}^{T_K^\epsilon} c_\epsilon(s_t,a_t) - KV(s_{\text{init}};\theta_*)\right] \\ &= \mathbb{E}\left[\sum_{t=1}^{T_K^\epsilon} c_\epsilon(s_t,a_t) - KV_\epsilon(s_{\text{init}};\theta_*)\right] + K\mathbb{E}\left[V_\epsilon(s_{\text{init}};\theta_*) - V(s_{\text{init}};\theta_*)\right]. \end{aligned} \tag{11}$$

Theorem 3.5 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 $B_\star^\epsilon \le B_\star + \epsilon T_\star$ (to see this note that $V_\epsilon(s;\theta_*) \le V_\epsilon^{\pi^*}(s;\theta_*) \le B_\star + \epsilon T_\star$). To bound the second term of (11), we have

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

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 $K \ge S^2A$ and $B_\star \le T_\star$ (since $c(s,a) \le 1$) 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. □

References

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.