Regret Bounds for Risk-Sensitive Reinforcement Learning under CVaR Objective

Presenter: Hao Liang
217019008@link.cuhk.edu.cn

The Chinese University of Hong Kong, Shenzhen, China

Oct. 27, 2022

Original paper: Bastani, O., Ma, Y. J., Shen, E., & Xu, W. (2022). Regret Bounds for Risk-Sensitive Reinforcement Learning. arXiv preprint arXiv:2210.05650.

Risk-sensitive Reinforcement Learning

- Standard RL focus on maximizing the expected return
- Risk-sensitive RL (RSRL) replaces the mean objective with risk measure that accounts for variation in possible outcomes
- Conditional value-at-risk (CVaR) is popular risk measure
  - the average risk at tail distribution of returns

Regret Bounds for RSRL

- Current works only focus on the entropic risk measure
- Regret bounds for more general risk measures are left open
- [Keramati et al.'20] proposes optimistic exploration for CVaR, but without any regret bounds.

Risk Measure

- Objective

$$\Phi(\pi) = \int_0^1 F^{\dagger}_{Z^{(\pi)}}(\tau) dG(\tau),$$

where

  - $Z^{(\pi)}$ is the return of policy π
  - $F^{\dagger}_{Z^{(\pi)}}$ is its quantile function/inverse CDF
  - G is a weighting function over the quantiles

- Captures a broad range of risk measures
  - mean: $G(\tau) = \tau \implies \int_0^1 F^{\dagger}_{Z^{(\pi)}}(\tau) d\tau = \int x dF(x)$
  - CVaR: $G(\tau) = \min\{\tau/\alpha, 1\} \implies \int_0^{\alpha} F^{\dagger}_{Z^{(\pi)}}(\tau) d\tau/\alpha = \frac{1}{\alpha} \int_0^{F^{-1}(\alpha)} x dF(x)$
  - value at risk (VaR): $G(\tau) = \mathbb{I}(\tau \leq \alpha) \implies F^{\dagger}_{Z^{(\pi)}}(\alpha)$

Regret Bound

- Consider the episodic MDP with regret minimization
- Propose an algorithm based on upper confidence bound strategy with regret bound

$$\mathrm{regret}(\mathfrak{A}) = \tilde{O}\left(T^{\frac{3}{2}} \cdot L_G \cdot |\mathcal{S}| \cdot \sqrt{|\mathcal{S}||\mathcal{A}|K}\right),$$

  where

  - T is the length of a single episode
  - L_G is the Lipschitz constant for G
  - K is the number of episodes
  - $|\mathcal{S}|$: the number of states, $|\mathcal{A}|$: the number of actions

Markov Decision Process

- $\mathcal{M} = (\mathcal{S}, \mathcal{A}, D, P, \mathbb{P}, T)$
  - initial state distribution $D(s)$
  - transition probabilities $P\left(s' \mid s, a\right)$
  - reward measure $\mathbb{P}_{R(s,a)}$, assume reward $r \in [0,1]$
- A history is a sequence

$$\xi \in \mathcal{Z} = \bigcup_{t=1}^{T} \mathcal{Z}_t \quad \text{where} \quad \mathcal{Z}_t = (\mathcal{S} \times \mathcal{A} \times \mathbb{R})^{t-1} \times \mathcal{S}$$

- Consider stochastic, history-dependent policies $\pi_t\left(a_t \mid \xi_t\right)_{t \in [T]}$

Markov Decision Process

- For all $\tau \in [T]$
$$\xi_\tau = \left(\left(s_1, a_1, r_1\right), \ldots, \left(s_{\tau-1}, a_{\tau-1}, r_{\tau-1}\right), s_\tau\right).$$
- History $\Xi_t^{(\pi)}$ generated by π up to step t
$$\mathbb{P}_{\Xi_t^{(\pi)}}\left(\xi_t\right) = \begin{cases} D\left(s_1\right) & \text{if } t = 1 \\ \mathbb{P}_{\Xi_{t-1}^{(\pi)}}\left(\xi_{t-1}\right) \cdot \pi_t\left(a_t \mid \xi_{t-1}\right) \cdot \mathbb{P}_{R(s_t, a_t)}\left(r_t\right) \cdot P\left(s_{t+1} \mid s_t, a_t\right) & \text{otherwise} \end{cases}$$
- An episode/rollout is a history $\xi \in \mathcal{Z}_T$ of length T generated by a given policy π.

Distributional Bellman Equation

- The return of π on step t is the r.v.

$$Z_t^{(\pi)}(\xi_t) = \sum_{\tau=t}^{T} r_\tau \mid \Xi_t^{(\pi)} = \xi_t$$

- Define $Z_{T+1}^{(\pi)}(\xi, s) = 0$, the distributional Bellman equation in the form of r.v.

$$Z_t^{(\pi)}(\xi) = R(s, a) + Z_{t+1}^{(\pi)}\left(\xi \circ (a, r, s')\right), a \sim \pi_t(\cdot \mid \xi), r \sim \mathbb{P}_{R(s,a)}, s' \sim P\left(\cdot \mid S(\xi), a\right)$$

- In the form of CDF

$$F_{Z_t^{(\pi)}(\xi)}(x) = \sum_{a \in \mathcal{A}} \pi_t(a \mid \xi) \sum_{s' \in \mathcal{S}} P\left(s' \mid S(\xi), a\right) \int F_{Z_{t+1}^{(\pi)}(\xi \circ (a, r, s'))}(x - r) \cdot dF_{R(s,a)}(r),$$

where $S(\xi) = s$ for $\xi = (\ldots, s)$ is the current state in history ξ

- The return of π is $Z^{(\pi)} = Z_1^{(\pi)}(s), s \sim D$

$$F_{Z^{(\pi)}}(\cdot) = \int F_{Z_1^{(\pi)}(s)}(\cdot) \cdot dD(s)$$

Risk-sensitive objective

- The quantile function of a r.v. X is

$$F_X^\dagger(\tau) = \inf\left\{x \in \mathbb{R} \mid F_X(x) \geq \tau\right\}, \tau \in [0,1]$$

- The risk-sensitive objective

$$\Phi_{\mathcal{M}}(\pi) = \int_0^1 F^\dagger_{Z^{(\pi)}}(\tau) \cdot dG(\tau)$$

- Optimal policy

$$\pi^*_{\mathcal{M}} \in \arg\max_\pi \Phi_{\mathcal{M}}(\pi)$$

Optimal Risk-Sensitive Policies

- There exists an optimal policy $\pi_t^* (a_t \mid y_t, s_t)$ that only depends on s_t and cumulative reward

$$y_t = \sum_{\tau=1}^{t-1} r_\tau$$

- Consider the augmented MDP $\tilde{\mathcal{M}} = (\tilde{\mathcal{S}}, \mathcal{A}, \tilde{D}, \tilde{P}, \tilde{\mathbb{P}}, T)$
  - $\tilde{\mathcal{S}} = \mathcal{S} \times \mathbb{R}$
  - $\tilde{D}((s, y)) = D(s) \cdot \delta_0(y)$
  - $\tilde{P}((s', y') \mid (s, y), a) = P(s' \mid s, a) \cdot \mathbb{P}_{R(s,a)}(y' - y)$
  - the rewards are now only provided on the final step

$$\mathbb{P}_{R_t((s,y),a)}(r) = \begin{cases} \delta_y(r) & \text{if } t = T \\ 0 & \text{otherwise} \end{cases}$$

Technical Assumptions

- **Assumption 1.** $F^{\dagger}_{Z^{(\pi)}}(1) = T \Leftrightarrow \mathbb{P}(Z^{(\pi)} = T) > 0.$
  the maximum reward is attained with some nontrivial probability.
- **Assumption 2.** G is L_G-Lipschitz continuous for some $L_G \in \mathbb{R}_{>0}$, and $G(0) = 0$.
  $L_G = \frac{1}{\alpha}$ for CVaR
- **Assumption 3.** We are given an algorithm for computing $\pi^*_{\mathcal{M}}$ for a given MDP $\mathcal{M}$.

Regret

- At the beginning of each episode $k \in [K]$, algorithm $\mathfrak{A}$ chooses a policy $\pi^{(k)} = \mathfrak{A}\left(H_k\right)$
- $H_k = \{\xi_{T,\kappa}\}_{\kappa=1}^{k-1}$ is the set of episodes observed so far
- Expected regret

$$\text{regret}(\mathfrak{A}) = \mathbb{E}\left[\sum_{k \in [K]} \Phi\left(\pi^*\right) - \Phi\left(\pi^{(k)}\right)\right]$$

- Assume that the initial state distribution D is known

Upper Confidence Bound Algorithm

- Construct an optimistic MDP $\mathcal{M}^{(k)}$ based on the history H_k
- Plan in $\mathcal{M}^{(k)}$ to obtain an optimistic policy $\pi^{(k)} = \pi^*_{\mathcal{M}^{(k)}}$
- Uses $\pi^{(k)}$ to act in the MDP for episode k

Algorithm 1 `Upper Confidence Bound Algorithm`

1: **for** $k \in [K]$ **do**
2: Compute $\mathcal{M}^{(k)}$ using prior episodes $\{\xi^{(i)} \mid i \in [k-1]\}$ and $\pi^{(k)} = \pi^*_{\mathcal{M}^{(k)}}$
3: Execute $\pi^{(k)}$ in the true MDP $\mathcal{M}$ and observe episode $\xi^{(k)}$
4: **end for**

Optimistic MDP

- Let $\tilde{\mathcal{M}}^{(k)}$ be the MDP using the empirical estimates $\tilde{P}^{(k)}$ and $F_{\tilde{R}^{(k)}}$
- Assume a distinguished state s_∞ with reward 1

$$P\left(s_\infty \mid s, a\right) = \mathbb{I}\left(s = s_\infty\right) \text{ and } P\left(s' \mid s_\infty, a\right) = \mathbb{I}\left(s' = s_\infty\right)$$

- Construction of $\hat{\mathcal{M}}^{(k)}$ uses s_∞ for optimism

$$\hat{P}^{(k)}\left(s' \mid s, a\right) = \begin{cases} \mathbb{I}\left(s' = s_\infty\right) & \text{if } s = s_\infty \\ 1 - \sum_{s' \in \mathcal{S} \setminus \{s_\infty\}} \tilde{P}^{(k)}\left(s' \mid s, a\right) & \text{if } s' = s_\infty \\ \max\left\{\tilde{P}^{(k)}\left(s' \mid s, a\right) - \epsilon_R^{(k)}(s, a), 0\right\} & \text{otherwise} \end{cases}$$

$$F_{\hat{R}^{(k)}(s,a)}(r) = \begin{cases} \mathbb{I}(r \geq 1) & \text{if } s = s_\infty \\ 1 & \text{if } r \geq 1 \\ \max\left\{F_{\tilde{R}^{(k)}(s,a)}(r) - \epsilon_R^{(k)}(s, a), 0\right\} & \text{otherwise} \end{cases}$$

Regret Upper Bound

Theorem 1.
For any $\delta \in (0, 1]$, *with probability at least* $1 - \delta$, *we have*

$$\text{regret}(\mathfrak{A}) \leq 4T^{3/2} \cdot L_G \cdot |\mathcal{S}| \cdot \sqrt{5|\mathcal{S}| \cdot |\mathcal{A}| \cdot K \cdot \log\left(\frac{4|\mathcal{S}| \cdot |\mathcal{A}| \cdot K}{\delta}\right)}$$

- $\tilde{\mathcal{O}}\left(T\sqrt{SAK}\right)$ improved bound of for UCBVI algorithm
- dependence on K is tight
- extra S factor
- dependence on A is tight
- extra $\sqrt{T}$ factor

Step 1: rewrite the objective

Lemma 2.

$$\Phi(\pi) = T - \int_{\mathbb{R}} G\left(F_{Z(\pi)}(x)\right) dx$$

Proof.

Use integration by parts

$$\begin{aligned}\Phi(\pi) &= \int_0^1 F^{\dagger}_{Z(\pi)}(\tau) \cdot dG(\tau) = \left[F^{\dagger}_{Z(\pi)}(\tau) \cdot G(\tau)\right]_0^1 - \int_0^1 G(\tau) \cdot dF^{\dagger}_{Z(\pi)}(\tau) \\ &= T - \int_0^1 G(\tau) \cdot dF^{\dagger}_{Z(\pi)}(\tau) = T - \int_{\mathbb{R}} G\left(F_{Z(\pi)}(x)\right) dx.\end{aligned}$$

■

Step 2: high prob. event

Given $\delta \in \mathbb{R}_{>0}$, define $\mathcal{E}$ to be the event where the following hold:

$$\left\|\tilde{P}^{(k)}(\cdot \mid s,a) - P(\cdot \mid s,a)\right\|_1 \leq \sqrt{\frac{2|\mathcal{S}|}{N^{(k)}(s,a)} \log\left(\frac{6|\mathcal{S}|\cdot|\mathcal{A}|\cdot K}{\delta}\right)} =: \epsilon_P^{(k)}(s,a) \quad (\forall s \in \mathcal{S}, a \in \mathcal{A})$$

$$\left\|F_{\tilde{R}(k)(s,a)} - F_{R(s,a)}\right\|_\infty \leq \sqrt{\frac{1}{2N^{(k)}(s,a)} \log\left(\frac{6|\mathcal{S}|\cdot|\mathcal{A}|\cdot K}{\delta}\right)} =: \epsilon_R^{(k)}(s,a) \quad (\forall s \in \mathcal{S}, a \in \mathcal{A})$$

$$\left\|\tilde{P}^{(k)}(\cdot \mid s,a) - P(\cdot \mid s,a)\right\|_\infty \leq \sqrt{\frac{1}{2N^{(k)}(s,a)} \log\left(\frac{6|\mathcal{S}|\cdot|\mathcal{A}|\cdot K}{\delta}\right)} = \epsilon_R^{(k)}(s,a) \quad (\forall s \in \mathcal{S}, a \in \mathcal{A}).$$

Lemma 3.

$\mathbb{P}\left[\mathcal{E}\right] \geq 1 - \delta..$

Step 3: Bound the objective difference

Lemma 4.
Consider $\mathcal{M} = (\mathcal{S}, \mathcal{A}, D, P, \mathbb{P}, T)$ *and* $\mathcal{M}' = (\mathcal{S}, \mathcal{A}, D, P', \mathbb{P}', T)$*, such that* $\|P'(\cdot \mid s, a) - P(\cdot \mid s, a)\|_1 \leq \epsilon_P(s, a)$ *and* $\left\|F_{R'(s,a)} - F_{R(s,a)}\right\|_\infty \leq \epsilon_R(s, a)$*. Then, we have*

$$|\Phi'(\pi) - \Phi(\pi)| \leq T \cdot L_G \cdot B(\pi) \quad (\forall k \in [K], \pi),$$

where

$$B(\pi) = \mathbb{E}_{\Xi_T^{(\pi)}}\left[\sum_{t=1}^{T} \epsilon_P(s_t, a_t) + \epsilon_R(s_t, a_t)\right].$$

Note that the expectation is taken w.r.t. the whole trajectory.

Proof of Lemma 4

Since $\Phi(\pi) = T - \int_{\mathbb{R}} G\left(F_{Z^{(\pi)}}(x)\right) dx$ by Lemma 2,

$$
\begin{aligned}
\left|\Phi'(\pi) - \Phi(\pi)\right| &= \left| \int_0^T \left(G\left(F_{Z'(\pi)}(x)\right) - G\left(F_{Z^{(\pi)}}(x)\right)\right) \cdot dx \right| \\
&\leq L_G \int_0^T \left| F_{Z'(\pi)}(x) - F_{Z^{(\pi)}}(x) \right| dx \\
&\leq L_G \cdot T \cdot \sup_x \left| F_{Z'(\pi)}(x) - F_{Z^{(\pi)}}(x) \right| \\
&= T \cdot L_G \cdot \left\| F_{Z'(\pi)}(x) - F_{Z^{(\pi)}}(x) \right\|_\infty .
\end{aligned}
$$

It suffices to show

$$
\left\| F_{Z'(\pi)} - F_{Z^{(\pi)}} \right\|_\infty \leq B(\pi) = \mathbb{E}_{\Xi_T^{(\pi)}} \left[\sum_{t=1}^{T} \epsilon_P\left(s_t, a_t\right) + \epsilon_R\left(s_t, a_t\right) \right].
$$

Proof of Lemma 4

$$\sup_{x\in\mathbb{R}}\left|F_{Z_t'^{(\pi)}(s,y)}(x)-F_{Z_t^{(\pi)}(s,y)}(x)\right|$$

$$\leq \sup_{x\in\mathbb{R}} \Big| \sum_{a\in\mathcal{A}}\sum_{s'\in\mathcal{S}}\pi(a\mid s,y)\left(P'\left(s'\mid s,a\right)-P\left(s'\mid s,a\right)\right)\int F_{Z_{t+1}'^{(\pi)}(s',y+r)}(x-r)dF_{R'(s,a)}(r)$$

$$+\sum_{a\in\mathcal{A}}\sum_{s'\in\mathcal{S}}\pi(a\mid s,y)P\left(s'\mid s,a\right)\int\left(F_{Z_{t+1}'^{(\pi)}(s',y+r)}(x-r)-F_{Z_{t+1}^{(\pi)}(s',y+r)}(x-r)\right)dF_{R'(s,a)}(r)$$

$$-\sum_{a\in\mathcal{A}}\sum_{s'\in\mathcal{S}}\pi(a\mid s,y)P\left(s'\mid s,a\right)\int\left(F_{R'(s,a)}(r)-F_{R(s,a)}(r)\right)dF_{Z_{t+1}^{(\pi)}(s',y+r)}(x-r)\Big|$$

$$\leq \sum_{a\in\mathcal{A}}\sum_{s'\in\mathcal{S}}\pi(a\mid s,y)\cdot\left|P'\left(s'\mid s,a\right)-P\left(s'\mid s,a\right)\right|$$

$$+\sum_{a\in\mathcal{A}}\sum_{s'\in\mathcal{S}}\pi(a\mid s,y)P(s'|s,a)\cdot\int\sup_{x'\in\mathbb{R}}\left|F_{Z_{t+1}'^{(\pi)}(s',y+r)}\left(x'\right)-F_{Z_{t+1}^{(\pi)}(s',y+r)}\left(x'\right)\right|\cdot d\mathbb{P}'_{R'(s,a)}(r)$$

$$+\sum_{a\in\mathcal{A}}\pi(a\mid s,y)\cdot\sup_{r'\in\mathbb{R}}\left|F_{R'(s,a)}\left(r'\right)-F_{R(s,a)}\left(r'\right)\right|$$

$$\leq \mathbb{E}\left[\epsilon_P(s,a)+\epsilon_R(s,a)\right]+\mathbb{E}\left[\sup_{x'\in\mathbb{R}}\left|F_{Z_{t+1}'^{(\pi)}(s',y+r)}\left(x'\right)-F_{Z_{t+1}^{(\pi)}(s',y+r)}\left(x'\right)\right|\right]$$

Proof of Lemma 4

$$
\begin{aligned}
\epsilon_t^{(\pi)} &:= \mathbb{E}\left[\sup_{x\in\mathbb{R}}\left|F_{Z_t'^{(\pi)}(s,y)}(x) - F_{Z_t^{(\pi)}(s,y)}(x)\right|\right] \\
&\leq \mathbb{E}\left[\epsilon_P(s,a) + \epsilon_R(s,a) + \sup_{x'\in\mathbb{R}}\left|F_{Z_{t+1}'^{(\pi)}(s',y+r)}(x') - F_{Z_{t+1}^{(\pi)}(s',y+r)}(x')\right|\right] \\
&= \mathbb{E}\left[\epsilon_P(s,a) + \epsilon_R(s,a)\right] + \epsilon_{t+1}^{(\pi)} \\
&= \mathbb{E}\left[\sum_{\tau=t}^{T}\epsilon_P\left(s_\tau, a_\tau\right) + \epsilon_R\left(s_\tau, a_\tau\right)\right],
\end{aligned}
$$

Thus

$$
\begin{aligned}
\left\|F_{Z'(\pi)} - F_{Z(\pi)}\right\|_\infty &= \sup_{x\in\mathbb{R}}\left|\mathbb{E}\left[F_{Z_1'^{(\pi)}(s)}(x) - F_{Z_1^{(\pi)}(s)}(x)\right]\right| \leq \mathbb{E}\left[\sup_{x\in\mathbb{R}}\left|F_{Z_1'^{(\pi)}(s)}(x) - F_{Z_1^{(\pi)}(s)}(x)\right|\right] \\
&= \epsilon_1^{(\pi)} \leq \mathbb{E}\left[\sum_{\tau=1}^{T}\epsilon_P\left(s_\tau, a_\tau\right) + \epsilon_R\left(s_\tau, a_\tau\right)\right] = B(\pi).
\end{aligned}
$$

Step 3: Bound the objective difference

- let $\Phi = \Phi_{\mathcal{M}}$, $\tilde{\Phi}^{(k)} = \Phi_{\tilde{\mathcal{M}}^{(k)}}$, and $\hat{\Phi}^{(k)} = \Phi_{\hat{\mathcal{M}}^{(k)}}$
- let $\pi^* = \pi_{\mathcal{M}^*}, \tilde{\pi}^{(k)} = \pi^*_{\tilde{\mathcal{M}}^{(k)}}$, and $\hat{\pi}^{(k)} = \pi^*_{\hat{\mathcal{M}}^{(k)}}$

Lemma 5.

On event $\mathcal{E}$, for all $k \in [K]$ and any policy π, we have

$$\left|\hat{\Phi}^{(k)}(\pi) - \Phi(\pi)\right| \leq 2T \cdot L_G \cdot \sqrt{|\mathcal{S}|} \cdot B^{(k)}(\pi),$$

where

$$B^{(k)}(\pi) = \mathbb{E}_{\Xi_T^{(\pi)}}\left[\sum_{t=1}^{T} \epsilon_P^{(k)}(s_t, a_t) + \epsilon_R^{(k)}(s_t, a_t)\right].$$

Proof of Lemma 5

$\hat{\mathcal{M}}^{(k)}$ and $\mathcal{M}$ satisfies that

$$\left\|\hat{P}^k(s,a) - P(s,a)\right\|_1 \leq \left\|\hat{P}^k(s,a) - \tilde{P}^k(s,a)\right\|_1 + \left\|\tilde{P}^k(s,a) - P(s,a)\right\|_1$$
$$\leq 2S \cdot \epsilon_R^k(s,a) \leq 2\sqrt{S}\epsilon_P^k(s,a)$$

$$\left\|F_{\tilde{R}^k(s,a)} - F_{R(s,a)}\right\|_\infty \leq \left\|F_{\tilde{R}^k(s,a)} - F_{\hat{R}^k(s,a)}\right\|_\infty + \left\|F_{\hat{R}^k(s,a)} - F_{R(s,a)}\right\|_\infty \leq 2\epsilon_R^k(s,a)$$

Replace $\epsilon_P(s,a)$ by $2\sqrt{S}\epsilon_P^k(s,a)$, and $\epsilon_R(s,a)$ by $2\epsilon_R^k(s,a)$ in Lemma 4

$$\left\|F_{\hat{Z}(\pi)} - F_{Z(\pi)}\right\|_\infty \leq \mathbb{E}\left[\sum_{\tau=1}^{T} 2\sqrt{S}\epsilon_P^k(s,a) + 2\epsilon_R^k(s,a)\right]$$
$$\leq 2\sqrt{S}\mathbb{E}\left[\sum_{\tau=1}^{T} \epsilon_P^k(s,a) + \epsilon_R^k(s,a)\right] = 2\sqrt{S}B^{(k)}(\pi).$$

Step 4: Optimism

Lemma 6.
On event $\mathcal{E}$, we have $\hat{\Phi}^{(k)}(\pi) \geq \Phi(\pi)$ for all $k \in [K]$ and all policies π.

Final Proof

Conditioned on $\mathcal{E}$

$$\operatorname{regret}(\mathfrak{A}) = \sum_{k=1}^{K} \Phi\left(\pi^{*}\right) - \Phi\left(\hat{\pi}^{(k)}\right) \leq \sum_{k=1}^{K} \hat{\Phi}^{(k)}\left(\pi^{*}\right) - \Phi\left(\hat{\pi}^{(k)}\right)$$

$$\leq \sum_{k=1}^{K} \hat{\Phi}^{(k)}\left(\hat{\pi}^{(k)}\right) - \Phi\left(\hat{\pi}^{(k)}\right) \leq \sum_{k=1}^{K} 2T \cdot L_G \cdot \sqrt{|\mathcal{S}|} \cdot B^{(k)}\left(\hat{\pi}^{(k)}\right)$$

$$= 2TL_G \sqrt{5|\mathcal{S}|^2 \log\left(\frac{4|\mathcal{S}| \cdot |\mathcal{A}| \cdot K}{\delta}\right)} \cdot \mathbb{E}_{\Xi_T^{(\pi(1:K))}}\left[\sum_{k=1}^{K}\sum_{t=1}^{T} \frac{1}{\sqrt{N^{(k)}\left(s_t, a_t\right)}}\right]$$

$$\leq 2TL_G \sqrt{5|\mathcal{S}|^2 \log\left(\frac{4|\mathcal{S}| \cdot |\mathcal{A}| \cdot K}{\delta}\right)} \sqrt{2|\mathcal{S}| \cdot |\mathcal{A}| \cdot KT}.$$