No-Regret Shannon Entropy Regularized Neural Contextual Bandit Online Learning for Robotic Grasping: Supplementary Material

Kyungjae Lee, Jaegu Choy, Yunho Choi, Hogun Kee, and Songhwai Oh

APPENDIX

A. Infinite Exploration

Theorem 1. *For any arm a, the expected count has the following lower bound, $\mathbb{E}[N_a(t)] \geq ct$ where $c = \frac{1}{K}\exp(-\frac{1}{\alpha})$.*

Before starting the proof of Theorem 1, we first prove the following Lemma.

Lemma 1.1. *The policy of SERN has a constant lower bound greater than zero, i.e., $[\pi_t]_a \geq c > 0$, where $c = \frac{1}{K}\exp(-\frac{1}{\alpha})$.*

Proof of Lemma 1.1. For each round, the proposed method samples an action from

$$\pi_t := \arg\max_{\pi} \left\{ \mathop{\mathbb{E}}_{a\sim\pi} [\hat{r}_a(s_t;\theta_{t-1})] + \alpha S(\pi) \right\}.$$

Thus, the policy distribution is the optimal solution of

$$\max_{\pi} \left\{ \mathop{\mathbb{E}}_{a\sim\pi} [\hat{r}_a(s_t;\theta_{t-1})] + \alpha S(\pi) \right\}$$

which is a concave maximization problem since $\mathbb{E}_{a\sim\pi}[\hat{r}_a(s_t;\theta_{t-1})]$ is linear for π and $\alpha S(\pi)$ is concave for π. The domain of this problem has two constraints, i.e., $\sum_a \pi_a - 1 = 0$ and $\pi_a \geq 0$. Since the problem is concave, strong duality holds and let us denote a dual variable for $\sum_a \pi_a - 1 = 0$ as μ and dual variable for positivity $\pi_a \geq 0$ as λ_a. Then, from Karush-Kuhn-Tucker (KKT) conditions, we have

$$\hat{r}_a(s_t;\theta_{t-1}) - \alpha\ln(\pi_a) - \alpha + \lambda_a + \mu = 0.$$

We first compute μ by multiplying π_a to both sides and summing up with respect to a. Then, $\mu = \alpha - \alpha S(\pi) - \mathbb{E}_{a\sim\pi}[\hat{r}_a(s_t;\theta_{t-1})]$ where $\lambda_a\pi_a = 0$, one of KKT conditions, is used. By using $S(\pi) \leq -\ln(1/K)$ and $\mathbb{E}_{a\sim\pi}[\hat{r}_a(s_t;\theta_{t-1})] \leq 1$, $\mu \geq \alpha + \alpha\ln(1/K) - 1$. Since $\ln(x)$ requires $x > 0$ and for all a, $\pi_a > 0$ holds, $\lambda_a = 0$ for all a from KKT conditions. Thus,

$$\ln(\pi_a) = \frac{\hat{r}_a(s_t;\theta_{t-1}) - \alpha + \mu}{\alpha} \geq \ln(1/K) - \frac{1}{\alpha}$$

where $\hat{r}_a \geq 0$. Finally, we get

$$\pi_a \geq \frac{1}{K}\exp\left(-\frac{1}{\alpha}\right).$$

□

The proof of Theorem 1 is as follows.

Proof of Theorem 1. Using Lemma 1.1, for all t and a, $[\pi_t]_a \geq c$ where $c = \frac{1}{K}\exp(-\frac{1}{\alpha})$. Thus, $\mathbb{E}[N_a(t)] = \sum_t [\pi_t]_a \geq ct$. □

Theorem 2. *For any arm a, let $N'_t := N_a(t) - ct$. Then, N'_t is submartingale and, from this fact, the following inequality holds, for any $\delta > 0$,*

$$\mathbb{P}(N_a(t) < ct - \delta) \leq \exp\left(-\frac{\delta^2}{8t}\right).$$

Proof of Theorem 2. Let $N'_a(t) = N_a(t) - ct$. To prove that $N'_a(t)$ is sub-Martingale, we need to check $\mathbb{E}[N'_a(t)|N'_a(t-1)] \geq N'_a(t-1)$. The inequality holds as follows:

$$\begin{aligned}
&\mathbb{E}[N'_a(t)|N'_a(t-1)] = \mathbb{E}[N_a(t) - ct|N'_a(t-1)] \\
&\quad= \mathbb{E}[N_a(t-1) - c(t-1) + \mathbb{I}(a_t = a) - c|N'_a(t-1)] \\
&\quad= N'_a(t-1) + \mathbb{E}[\mathbb{I}(a_t = a) - c|N'_a(t-1)] \\
&\quad= N'_a(t-1) + [\pi_t]_a - c \\
&\quad\geq N'_a(t-1) \;\; (\because [\pi_t]_a \geq c).
\end{aligned}$$

For sub-Martingale random variable, since $|N'_a(t) - N'_a(t-1)| < 1 + c < 2$ for all t, Azuma-Hoeffding inequality holds, $\mathbb{P}(N'_a(t) - N'_a(0) \leq -\delta) = \mathbb{P}(N_a(t) \leq ct - \delta) \leq \exp\left(-\frac{\delta^2}{8t}\right)$. □

B. Regret Bound

Theorem 3. *For $\alpha > 0$ and $1 > q > 0$, the expected cumulative regret of SERN is bounded as*

$$\begin{aligned}
\mathcal{R}_T \leq& \beta\sum_{t=1}^{T} \mathop{\mathbb{E}}_{s_{1:t},a_{1:t}}\left[\frac{1}{\sqrt{(N_{a^\star}(t-1)+1)}}\right] \\
&+ \beta\sum_{t=1}^{T} \mathop{\mathbb{E}}_{s_{1:t},a_{1:t}}\left[\frac{1}{\sqrt{(N_{a_t}(t-1)+1)}}\right] \\
&+ \sum_{t=1}^{T}\mathbb{P}(a^\star \neq \hat{a}^\star_{t-1}) + \alpha\ln(K)T,
\end{aligned}$$

where $K = |\mathcal{A}|$, $a^\star = \arg\max_a \mathbb{E}_s[r_a(s)]$, and $\hat{a}^\star_t = \arg\max_a \mathbb{E}_s[\hat{r}_a(s;\theta_t)]$.

Before proving the regret bound, we introduce a new lemma for our policy distribution.

Lemma 3.1. *For any vector $r \in \mathbb{R}^{|\mathcal{A}|}$, let a distribution be $\pi := \arg\max_{\pi'}\{\mathbb{E}_{a\sim\pi'}[r_a] + \alpha S(\pi')\}$. Then,*

$$\max_a r_a - \mathbb{E}_{a\sim\pi}[r_a] \leq \alpha\ln(K)$$

K. Lee, J. Choy, Y. Choi, H. Kee, and S. Oh are with the Department of Electrical and Computer Engineering and ASRI, Seoul National University, Seoul 08826, Korea (e-mail: {kyungjae.lee, jaegu.choy, yunho.choi}@rllab.snu.ac.kr, {hogunkee, songhwai}@snu.ac.kr)

where $K = |\mathcal{A}|$

Proof of Lemma 3.1. Let $\pi'' := \arg\max_{\pi'} \mathbb{E}_{a\sim\pi'}[r_a]$, Then,

$$\begin{aligned}\max_a r_a &= \mathop{\mathbb{E}}_{a\sim\pi''}[r_a] = \mathop{\mathbb{E}}_{a\sim\pi''}[r_a] + \alpha S(\pi'') \ \ (\because S(\pi'')=0)\\ &\le \mathop{\mathbb{E}}_{a\sim\pi}[r_a] + \alpha S(\pi) \le \mathop{\mathbb{E}}_{a\sim\pi}[r_a] + \alpha \max_{\pi'} S(\pi')\\ &= \mathop{\mathbb{E}}_{a\sim\pi}[r_a] + \alpha\ln(K)\end{aligned}$$

Consequently, $\max_a r_a - \mathbb{E}_{a\sim\pi}[r_a] \le \alpha\ln(K)$ □

By using this Lemma, we prove the Theorem 3.

Proof of Theorem 3.

$$\begin{aligned}\mathcal{R}_T &= \sum_{t=1}^{T} \max_{a'} \mathop{\mathbb{E}}_{s_{1:T}}[r_{a'}(s_t)] - \mathop{\mathbb{E}}_{s_{1:T},a_{1:T}}[r_{a_t}(s_t)]\\ &\le \sum_{t=1}^{T} \max_{a'} \mathop{\mathbb{E}}_{s_t}[r_{a'}(s_t)] - \mathop{\mathbb{E}}_{s_t,a_{1:t}}[r_{a_t}(s_t)].\end{aligned}$$

We first compute the bound of the regret for each round $\max_{a'} \mathbb{E}_{s_t}[r_{a'}(s_t)] - \mathbb{E}_{s_t,a_{1:t}}[r_{a_t}(s_t)]$.

Let us define $a^\star := \arg\max_{a'} \mathbb{E}_s[r_{a'}(s)]$ and $\hat{a}^\star_{t-1} := \arg\max_{a'} \mathbb{E}_s[\hat{r}_{a'}(s;\theta_{t-1})]$. Then, the regret at round t is

$$\begin{aligned}&\max_{a'} \mathop{\mathbb{E}}_{s_t}[r_{a'}(s_t)] - \mathop{\mathbb{E}}_{s_t,a_{1:t}}[r_{a_t}(s_t)]\\ &= \mathop{\mathbb{E}}_{s_t}[r_{a^\star}(s_t)] - \mathop{\mathbb{E}}_{s_{1:t},a_{1:t}}[\hat{r}_{a^\star}(s_t;\theta_{t-1})] \quad (1)\\ &+ \mathop{\mathbb{E}}_{s_{1:t},a_{1:t}}[\hat{r}_{a^\star}(s_t;\theta_{t-1})] - \mathop{\mathbb{E}}_{s_{1:t},a_{1:t}}\left[\hat{r}_{\hat{a}^\star_{t-1}}(s_t;\theta_{t-1})\right] \quad (2)\\ &+ \mathop{\mathbb{E}}_{s_{1:t},a_{1:t}}\left[\hat{r}_{\hat{a}^\star_{t-1}}(s_t;\theta_{t-1})\right] - \mathop{\mathbb{E}}_{s_{1:t},a_{1:t}}[\hat{r}_{a_t}(s_t;\theta_{t-1})] \quad (3)\\ &+ \mathop{\mathbb{E}}_{s_{1:t},a_{1:t}}[\hat{r}_{a_t}(s_t;\theta_{t-1})] - \mathop{\mathbb{E}}_{s_t,a_{1:t}}[r_{a_t}(s_t)]. \quad (4)\end{aligned}$$

From Assumption 3, the (1) and (4) terms are caused by an estimation error and are bounded as follows:

$$\begin{aligned}&\mathop{\mathbb{E}}_{s_{1:t},a_{1:t}}[\hat{r}_{a_t}(s_t;\theta_{t-1}) - r_{a_t}(s_t;\theta_{t-1})]\\ &\quad\le \mathop{\mathbb{E}}_{s_{1:t},a_{1:t}}[|\hat{r}_{a_t}(s_t;\theta_{t-1}) - r_{a_t}(s_t;\theta_{t-1})|]\\ &\quad\le \beta \mathop{\mathbb{E}}_{s_{1:t},a_{1:t}}\left[\frac{1}{\sqrt{(N_{a_t}(t-1)+1)}}\right]\end{aligned}$$

and, similarly,

$$\begin{aligned}&\mathop{\mathbb{E}}_{s_{1:t},a_{1:t}}[\hat{r}_{a^\star}(s_t;\theta_{t-1}) - r_{a^\star}(s_t;\theta_{t-1})]\\ &\quad\le \beta \mathop{\mathbb{E}}_{s_{1:t},a_{1:t}}\left[\frac{1}{\sqrt{(N_{a^\star}(t-1)+1)}}\right].\end{aligned}$$

the (2) term comes from the failure probability for classifying the optimal action using $\hat{r}_a(s_t)$. Thus, we can rewrite it as follows:

$$\begin{aligned}&\mathop{\mathbb{E}}_{s_{1:t},a_{1:t}}[\hat{r}_{a^\star}(s_t;\theta_{t-1})] - \mathop{\mathbb{E}}_{s_{1:t},a_{1:t}}\left[\hat{r}_{\hat{a}^\star_{t-1}}(s_t;\theta_{t-1})\right]\\ &= \mathop{\mathbb{E}}_{s_{1:t},a_{1:t}}\left[\mathbb{I}(a^\star \ne \hat{a}^\star_{t-1})(\hat{r}_{a^\star}(s_t;\theta_{t-1}) - \hat{r}_{\hat{a}^\star_{t-1}}(s_t;\theta_{t-1}))\right]\\ &\le \mathop{\mathbb{E}}_{s_{1:t},a_{1:t}}[\mathbb{I}(a^\star \ne \hat{a}^\star_{t-1})] = \mathbb{P}(a^\star \ne \hat{a}^\star_{t-1}).\end{aligned}$$

The (3) term is bounded by Lemma 3.1,

$$\begin{aligned}&\mathop{\mathbb{E}}_{s_{1:t},a_{1:t}}\left[\hat{r}_{\hat{a}^\star_{t-1}}(s_t;\theta_{t-1})\right] - \mathop{\mathbb{E}}_{s_{1:t},a_{1:t}}[\hat{r}_{a_t}(s_t;\theta_{t-1})]\\ &\le \max_a \mathop{\mathbb{E}}_{s_{1:t},a_{1:t}}[\hat{r}_a(s_t;\theta_{t-1})] - \mathop{\mathbb{E}}_{a_t\sim\pi_t}\mathop{\mathbb{E}}_{s_{1:t},a_{1:t-1}}[\hat{r}_{a_t}(s_t;\theta_{t-1})]\\ &\le \alpha\ln(K)\end{aligned}$$

Finally, we have,

$$\begin{aligned}&\max_{a'} \mathop{\mathbb{E}}_{s_t}[r_{a'}(s_t)] - \mathop{\mathbb{E}}_{s_t,a_{1:t}}[r_{a_t}(s_t)]\\ &\le \beta \mathop{\mathbb{E}}_{s_{1:t},a_{1:t}}\left[\frac{1}{\sqrt{(N_{a^\star}(t-1)+1)}}\right]\\ &\quad + \beta \mathop{\mathbb{E}}_{s_{1:t},a_{1:t}}\left[\frac{1}{\sqrt{(N_{a_t}(t-1)+1)}}\right]\\ &\quad + \mathbb{P}(a^\star \ne \hat{a}^\star_{t-1}) + \alpha\ln(K).\end{aligned}$$

Consequently, for the expected cumulative regret,

$$\begin{aligned}\mathcal{R}_T \le&\beta\sum_{t=1}^{T} \mathop{\mathbb{E}}_{s_{1:t},a_{1:t}}\left[\frac{1}{\sqrt{(N_{a^\star}(t-1)+1)}}\right]\\ &+ \beta\sum_{t=1}^{T} \mathop{\mathbb{E}}_{s_{1:t},a_{1:t}}\left[\frac{1}{\sqrt{(N_{a_t}(t-1)+1)}}\right]\\ &+ \sum_{t=1}^{T} \mathbb{P}(a^\star \ne \hat{a}^\star_{t-1}) + \alpha\ln(K)T.\end{aligned}$$

□

Theorem 4. *Let* $\alpha = \frac{\alpha_0}{\ln(T^p)}$ *for* $\alpha_0 > 0$*. Then, the expected cumulative regret of SERN is bounded as*

$$\begin{aligned}\mathcal{R}_T \le&\frac{C_0}{c_0^{3/2}} T^{\frac{3p+1}{2}} + C_1\left(1 - \exp\left(-c_0^2 d_1 T^{-2p}\right)\right)^{-1}\\ &+ C_2\left(1 - \exp\left(-c_0^2 d_2 T^{-2p}\right)\right)^{-1} + \alpha_0\ln(K)T\left(\ln(T^p)\right)^{-1},\end{aligned}$$

where $c_0 = \exp(-1/\alpha_0)$, $C_0 = 2^{\frac{7}{2}} K^{\frac{3}{2}}\beta$, $C_1 = 2\beta K$, $C_2 = 2(K-1)\exp((\beta/\Delta_2)^2 - 1/4)$, $d_1 = 1/(32K^2)$, *and* $d_2 = 1/(8K^2)$.

Proof of Theorem 4. From Theorem 3, it is known that the expected regret is bounded by three terms: estimation error, the failure probability, and regularization. For $\mathbb{E}_{s_{1:t},a_{1:t}}\left[\frac{1}{\sqrt{(N_a(t-1)+1)}}\right]$, since the proposed method explores every arms infinitely, estimation errors of all arms become zero. Now, for any a, we can compute the upper bound by using Theorem 1 and 2,

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{s_{1:t},a_{1:t}}\left[\frac{1}{\sqrt{(N_a(t-1)+1)}}\right] \\
&= \mathop{\mathbb{E}}_{s_{1:t},a_{1:t}}\left[\frac{1}{\sqrt{(N_a(t-1)+1)}}\mathbb{I}\left(N_a(t-1) > \frac{ct}{2}\right)\right] \\
&\quad + \mathop{\mathbb{E}}_{s_{1:t},a_{1:t}}\left[\frac{1}{\sqrt{(N_a(t-1)+1)}}\mathbb{I}\left(N_a(t-1) \leq \frac{ct}{2}\right)\right] \\
&\leq \mathop{\mathbb{E}}_{s_{1:t},a_{1:t}}\left[\sqrt{\frac{2}{ct}}\mathbb{I}\left(N_a(t-1) > \frac{ct}{2}\right)\right] \\
&\quad + \mathop{\mathbb{E}}_{s_{1:t},a_{1:t}}\left[\mathbb{I}\left(N_a(t-1) \leq \frac{ct}{2}\right)\right] \\
&\leq \sqrt{\frac{2}{ct}}\mathbb{P}\left(N_a(t-1) > \frac{ct}{2}\right) + \mathbb{P}\left(N_a(t-1) \leq \frac{ct}{2}\right) \\
&\leq \sqrt{\frac{2}{ct}} \cdot \frac{2\mathbb{E}\left[N_a(t-1)\right]}{ct} + \mathbb{P}\left(N_a(t-1) \leq \frac{ct}{2}\right) \\
&\leq \sqrt{\frac{2}{ct}} \cdot \frac{2t}{ct} + \mathbb{P}\left(N_a(t-1) \leq \frac{ct}{2}\right) \\
&\leq \frac{2^{3/2}}{c^{3/2}}\frac{1}{\sqrt{t}} + \exp\left(-\frac{c^2 t}{32}\right)
\end{aligned}
$$

where for the last inequality we use the Markov inequality and the Azuma Hoeffding inequality, respectively. Finally, we get

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{s_{1:t-1},a_{1:t}}\left[\frac{1}{\sqrt{(N_{a_t}(t-1)+1)}}\right] \\
&= \mathop{\mathbb{E}}_{a_t}\left[\mathop{\mathbb{E}}_{s_{1:t-1},a_{1:t-1}}\left[\frac{1}{\sqrt{(N_{a_t}(t-1)+1)}}\right]\right] \\
&\leq \mathop{\mathbb{E}}_{a_t}\left[\frac{2^{3/2}}{c^{3/2}}\frac{1}{\sqrt{t}} + \exp\left(-\frac{c^2 t}{32}\right)\right] \leq \frac{2^{3/2}}{c^{3/2}}\frac{1}{\sqrt{t}} + \exp\left(-\frac{c^2 t}{32}\right)
\end{aligned}
$$

and

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{s_{1:t-1},a_{1:t}}\left[\frac{1}{\sqrt{(N_{a^\star}(t-1)+1)}}\right] \\
&= \mathop{\mathbb{E}}_{a_t}\left[\mathop{\mathbb{E}}_{s_{1:t-1},a_{1:t-1}}\left[\frac{1}{\sqrt{(N_{a^\star}(t-1)+1)}}\right]\right] \\
&\leq \mathop{\mathbb{E}}_{a_t}\left[\frac{2^{3/2}}{c^{3/2}}\frac{1}{\sqrt{t}} + \exp\left(-\frac{c^2 t}{32}\right)\right] \leq \frac{2^{3/2}}{c^{3/2}}\frac{1}{\sqrt{t}} + \exp\left(-\frac{c^2 t}{32}\right).
\end{aligned}
$$

For the failure probability $\mathbb{P}(a^\star \neq \hat{a}^\star_{t-1})$, let us define an estimation error bound of Assumption 3 as $\beta_{N_a(t-1)} := \frac{\beta}{\sqrt{N_a(t-1)+1}}$. We obtain the bound as follows:

$$
\begin{aligned}
\mathbb{P}\left(a^\star \neq \hat{a}^\star_{t-1}\right) &= \mathbb{P}\left(\hat{r}_{a^\star}(s_t) < \hat{r}_{\hat{a}^\star_{t-1}}(s_t)\right) \\
&\leq \sum_{a \neq a^\star} \mathbb{P}\left(\hat{r}_{a^\star}(s_t) < \hat{r}_a(s_t)\right) \\
&\leq \sum_{a \neq a^\star} \mathbb{P}\left(r_{a^\star}(s_t) - \beta_{N_{a^\star}(t-1)} < r_a(s_t) + \beta_{N_a(t-1)}\right) \\
&\leq \sum_{a \neq a^\star} \mathbb{P}\left(\Delta_a(s_t) < \beta_{N_{a^\star}(t-1)} + \beta_{N_a(t-1)}\right) \\
&\leq \sum_{a \neq a^\star} \mathbb{P}\left(\Delta_2 < \beta_{N_{a^\star}(t-1)} + \beta_{N_a(t-1)}\right) \\
&\leq \sum_{a \neq a^\star} \mathbb{P}\left(\frac{\Delta_2}{2} < \beta_{N_{a^\star}(t-1)}\right) + \mathbb{P}\left(\frac{\Delta_2}{2} < \beta_{N_a(t-1)}\right).
\end{aligned}
$$

Now, we can bound $\mathbb{P}\left(\frac{\Delta_2}{2} < \beta_{N_a(t-1)}\right)$ using Theorem 2,

$$
\begin{aligned}
&\mathbb{P}\left(\frac{\Delta_2}{2} < \beta_{N_a(t-1)}\right) = \mathbb{P}\left(N_a(t-1) < \left(\frac{2\beta}{\Delta_2}\right)^2 - 1\right) \\
&\leq \exp\left(-\frac{(ct - (2\beta/\Delta_2)^2 + 1)^2}{8t}\right) \\
&= \exp\left(-\frac{c^2 t}{8} + \frac{(2\beta/\Delta_2)^2 - 1}{4} - \frac{((2\beta/\Delta_2)^2 - 1)^2}{8t}\right) \\
&\leq \exp\left(\frac{(2\beta/\Delta_2)^2 - 1}{4}\right)\exp\left(-\frac{c^2 t}{8}\right)
\end{aligned}
$$

Hence, we get,

$$
\begin{aligned}
&\mathbb{P}\left(a^\star \neq \hat{a}^\star_{t-1}\right) \\
&\quad \leq \sum_{a \neq a^\star} 2\exp\left(\frac{(2\beta/\Delta_2)^2 - 1}{4}\right)\exp\left(-\frac{c^2 t}{8}\right) \\
&\quad = 2(K-1)\exp\left((\beta/\Delta_2)^2 - 1/4\right)\exp\left(-\frac{c^2 t}{8}\right)
\end{aligned}
$$

Let $C_0 = 2^{\frac{7}{2}} K^{\frac{3}{2}} \beta$, $C_1 = 2\beta$, $C_2 = 2(K-1)\exp((\beta/\Delta_2)^2 - 1/4)$, $d_1 = 1/(32K^2)$, and $d_2 = 1/(8K^2)$. By combining all bounds, $\mathcal{R}_T$ can be bounded as follows:

$$
\begin{aligned}
\mathcal{R}_T \leq& \frac{2^{5/2}\beta}{c^{3/2}}\sum_{t=1}^{T}\frac{1}{\sqrt{t}} + 2\beta\sum_{t=1}^{T}\exp\left(-\frac{c^2 t}{32}\right) \\
&+ 2(K-1)\exp\left((\beta/\Delta_2)^2 - 1/4\right)\sum_{t=1}^{T}\exp\left(-\frac{c^2 t}{8}\right) \\
&+ \alpha\ln(K)T \\
=& \frac{C_0 K^{-3/2}/2}{c^{3/2}}\sum_{t=1}^{T}\frac{1}{\sqrt{t}} + C_1\sum_{t=1}^{T}\exp\left(-\frac{c^2 t}{32}\right) \\
&+ C_2\sum_{t=1}^{T}\exp\left(-\frac{c^2 t}{8}\right) + \alpha\ln(K)T \\
\leq& \frac{C_0 K^{-3/2}/2}{c^{3/2}}(1 + 2\sqrt{T} - 2\sqrt{2}) \\
&+ C_1\frac{\exp\left(-c^2 T/32\right) - 1}{\exp\left(-c^2/32\right) - 1} \\
&+ C_2\frac{\exp\left(-c^2 T/8\right) - 1}{\exp\left(-c^2/8\right) - 1} + \alpha\ln(K)T \\
\leq& \frac{C_0 K^{-3/2}}{c^{3/2}}\sqrt{T} + \frac{C_1}{1 - \exp\left(-c^2/32\right)} + \frac{C_2}{1 - \exp\left(-c^2/8\right)} \\
&+ \alpha\ln(K)T.
\end{aligned}
$$

Note that all terms are sub-linear except for $\alpha\ln(K)T$. To make $\alpha\ln(K)T$ sub-linear, we set α to be $\alpha_0(\ln(T^p))^{-1}$ with $\alpha_0 > 0$. Then, the lower bound c becomes $\frac{\exp\left(-\frac{1}{\alpha_0}\right)}{KT^p}$

and let $c_0 := \exp\left(-\frac{1}{\alpha_0}\right)$. Finally,

$$\begin{aligned}
\mathcal{R}_T \leq & \frac{C_0 K^{-3/2}}{c^{3/2}}\sqrt{T} + \frac{C_1}{1-\exp\left(-c^2/32\right)} + \frac{C_2}{1-\exp\left(-c^2/8\right)} \\
& + \alpha \ln(K) T \\
\leq & \frac{C_0}{c_0^{3/2}} T^{\frac{3p+1}{2}} + C_1(1-\exp(-T^{-2p}\cdot c_0^2/(32K^2)))^{-1} \\
& + C_2(1-\exp(-T^{-2p}\cdot c_0^2/(8K^2)))^{-1} \\
& + \alpha_0 \ln(K) T(\ln(T^p))^{-1} \\
\leq & \frac{C_0}{c_0^{3/2}} T^{\frac{3p+1}{2}} + C_1(1-\exp(-c_0^2 d_1 T^{-2p}))^{-1} \\
& + C_2(1-\exp(-c_0^2 d_2 T^{-2p}))^{-1} \\
& + \alpha_0 \ln(K) T(\ln(T^p))^{-1}.
\end{aligned}$$

□

Theorem 5. *For $1/3 > p > 0$, if the number of rounds, T, goes to infinity, then, time-averaged regret converges to zero:* $\lim_{T\to\infty} \frac{\mathcal{R}_T}{T} = 0$.

Proof of Theorem 5. To prove that $\lim_{T\to\infty} \frac{\mathcal{R}_T}{T} = 0$, we show that the upper bound of $\mathcal{R}_T/T$ converges to zero, then, proof will be done since the lower bound of $\mathcal{R}_T/T$ is also zero.

$$\begin{aligned}
\frac{\mathcal{R}_T}{T} \leq & \frac{C_0}{c_0^{3/2}} T^{\frac{3p-1}{2}} + C_1(1-\exp(-d_1 T^{-2p}))^{-1} T^{-1} \\
& + C_2(1-\exp(-d_2 T^{-2p}))^{-1} T^{-1} \\
& + \ln(K)(\ln(T^p))^{-1}.
\end{aligned}$$

Since $1/3 > p > 0$, $T_{(3p-1)/2}$ converges to zero and $\ln(T^p)^{-1}$ also converges to zero. To show that the second and third terms converge to zero, we prove that, for a positive a, $\lim_{x\to\infty}(1-\exp(-ax^{-2p})x)^{-1}x^{-1} = 0$ as follows:

$$\lim_{x\to\infty}(1-\exp(-ax^{-2p}))^{-1} a x^{-2p} \cdot x^{2p-1}/a = 1 \cdot 0 = 0$$

where $\lim_{z\to 0} \frac{z}{\exp(z)-1} = 1$ is used. □