Supplementary Material: Regional Multi-Armed Bandits

Zhiyang Wang, Ruida Zhou, Cong Shen
School of Information Science and Technology
University of Science and Technology of China
{wzy43, zrd127}@mail.ustc.edu.cn, congshen@ustc.edu.cn

1 Proof of Theorem 4

Proof. First we define some events to analyze the suboptimal selections of arms:

- Event $A(t)$: "choosing a suboptimal arm at time t";
- Event $B(t)$: "choosing a suboptimal group at time t";
- Event $C(t)$: "choosing group m^* that contains the best arm at time t".

The probabilities of these events satisfy:

$$\begin{aligned}\mathbb{P}(A(t)) &= \mathbb{P}(A(t)B(t)) + \mathbb{P}(A(t)C(t)) \\ &\leq \mathbb{P}(B(t)) + \mathbb{P}(A(t)|C(t)).\end{aligned}$$

In the following, we will separately study the two events $B(t)$ and $A(t)|C(t)$.

I. Analyze the regret caused by choosing a suboptimal group

The selection among groups follows a UCB principle. Similar to the argument in [6], at least one of the following three inequalities must be true:

$$\mu_{m^*}(\hat{\theta}_{m^*}(t)) + \psi_{m^*}^{-1}\left(\frac{\alpha_{m^*}\log(t)}{N_{m^*}(t-1)}\right) \leq \mu_{m^*}(\theta_{m^*}),$$

$$\mu_m(\hat{\theta}_m(t)) > \psi_m^{-1}\left(\frac{\alpha_m \log(t)}{N_m(t-1)}\right) + \mu_m(\theta_m),$$

$$N_m(t-1) < \frac{\alpha_m \log(t)}{\psi_m(\Delta_m/2)} = u_m(t).$$

Next we show that the gap between the estimated and the true group performance can be bounded by a function of the gap between the estimated and the true expected reward of the most played arm $\hat{k}_m$ in the group. The following inequalities are based on Assumption 1 and Proposition 2, 3.

$$\begin{aligned}|\mu_m(\widehat{\theta_m}) - \mu_m(\theta_m)| &\leq |\mu_{m,k}(\widehat{\theta_m}) - \mu_{m,k}(\theta_m)| \\ &\leq D_{2,m}|\hat{\theta}_m - \theta_m|^{\gamma_{2,m}} \\ &\leq D_{2,m}|\bar{D}_{1,m}|\hat{X}_{m,\hat{k}_m} - X_{m,\hat{k}_m}|^{\bar{\gamma}_{1,m}}|^{\gamma_{2,m}} \\ &\leq D_{2,m}\bar{D}_{1,m}^{\gamma_{2,m}}|\hat{X}_{m,\hat{k}_m} - X_{m,\hat{k}_m}|^{\bar{\gamma}_{1,m}\gamma_{2,m}} \\ &= \phi_m(|\hat{X}_{m,\hat{k}_m} - X_{m,\hat{k}_m}|) \qquad (6)\end{aligned}$$

Let $N_m(T)$ be the total number of times group $m \neq m^*$ is chosen before time T. We have:

$$\begin{aligned}&\mathbb{E}[N_m(T)] = \sum_{t=1}^{T}\mathbb{P}(m(t) = m) \\ &\leq u_m(T) + \sum_{t=u_m(T)+1}^{T}(\mathbb{P}((19)\text{ true}) + \mathbb{P}((20)\text{ true})) \qquad (7)\end{aligned}$$

and,

$$\begin{aligned}&\mathbb{P}((19)\text{ true}) \leq \\ &\sum_{s=1}^{t}\mathbb{P}\left(\left|\mu_m\left(\hat{\theta}_m(s)\right) - \mu_m(\theta_m)\right| > \psi_m^{-1}\left(\frac{\alpha_m\log(t)}{s}\right)\right) \\ &\leq \sum_{s=1}^{t}\mathbb{P}\Big(\phi_m^{-1}(\psi_m^{-1}(\frac{\alpha_m\log(t)}{s})) \leq \\ &\qquad |\hat{X}_{m,k}(s') - \mu_{m,k}(\theta_m)|\Big) \\ &\leq \sum_{s=1}^{t}\exp\left(-2\Big(\phi_m^{-1}(\psi_m^{-1}(\frac{\alpha_m\log(t)}{s}))\Big)^2\frac{s}{K_m}\right) \qquad (8) \\ &\leq \sum_{s=1}^{t}\exp\left(-2\alpha_m\frac{\log(t)}{s}\frac{s}{K_m}\right) \\ &\leq \sum_{s=1}^{t}\frac{1}{t^\alpha} \leq \frac{1}{t^{\alpha-1}},\end{aligned}$$

where inequality (8) is obtained by the Chernoff-Hoeffding inequality. The same upper bound holds for $\mathbb{P}((20)\text{ true})$. Therefore inequality (7) can be further written as:

$$\mathbb{E}[N_m(T)] \leq u_m(T) + \frac{2}{\alpha - 2},$$

with $\alpha = 2\alpha_m / K_m > 2$. Thus, the regret caused by choosing the wrong group $R_B(T)$ satisfies:

$$R_B(T) \leq \sum_{m \neq m^*} \left(\frac{\alpha_m \log(T)}{\psi_m(\Delta_m/2)} + \frac{2}{\alpha - 2} \right).$$

II. Analyze the regret by choosing a suboptimal arm in the optimal group

Assume that we have chosen the optimal group $N_{m^*}(t)$ times before time t and $m(t) = m^*$. We now consider the probability of choosing a suboptimal arm in the optimal group:

$$\begin{aligned}
&\mathbb{P}(k(t) \neq k^*) \leq \mathbb{P}(\delta \leq |\hat{\theta}_{m^*}(t) - \theta_{m^*}|) \\
&= \mathbb{P}(\delta \leq |\mu^{-1}_{m^*,\hat{k}_{m^*}}(\hat{X}_{m^*,\hat{k}_{m^*}}(t)) - \mu^{-1}_{m^*,\hat{k}_{m^*}}(X_{m^*,\hat{k}_{m^*}})|) \qquad (9) \\
&\leq \mathbb{P}(\delta \leq \bar{D}_{1,m^*} |\hat{X}_{m^*,\hat{k}_{m^*}}(t) - X_{m^*,\hat{k}_{m^*}}|^{\bar{\gamma}_1}) \\
&= \mathbb{P}\left(\left(\frac{\delta}{\bar{D}_{1,m^*}} \right)^{\gamma_{1,m^*}} \leq |\hat{X}_{m^*,\hat{k}_{m^*}}(t) - X_{m^*,\hat{k}_{m^*}}| \right) \\
&\leq 2\exp\left(-2 \left(\frac{\delta}{\bar{D}_{1,m^*}} \right)^{2\gamma_{1,m^*}} N_{m^*,\hat{k}_{m^*}} \right) \qquad (10) \\
&\leq 2\exp\left(-2 \left(\frac{\delta}{\bar{D}_{1,m^*}} \right)^{2\gamma_{1,m^*}} \frac{N_{m^*}(t)}{K_{m^*}} \right). \qquad (11)
\end{aligned}$$

Equality (9) is obtained by the definition of $\hat{\theta}_m(t)$ and inequality (10) is obtained by the Chernoff-Hoeffding bound. Inequality (11) is obtained by the fact that $N_{m^*,\hat{k}_{m^*}} \geq \frac{N_{m^*}(t)}{K_{m^*}}$ because $\hat{k}_{m^*}$ is the arm that has been selected most in the group.

We denote $R_C(T)$ as the regret caused by choosing a suboptimal arm in the optimal group up to time T. It satisfies:

$$\begin{aligned}
R_C(T) &\leq \sum_{t=1}^{T} 2\exp\left(-2 \left(\frac{\delta}{\bar{D}_{1,m^*}} \right)^{2\gamma_{1,m^*}} \frac{t}{K_{m^*}} \right) \\
&< \frac{2}{\exp\left(\frac{2}{K_{m^*}} \left(\frac{\delta}{\bar{D}_{1,m^*}} \right)^{2\gamma_{1,m^*}} \right) - 1},
\end{aligned}$$

which is a bounded regret.

Finally, the total regret can be derived as:

$$\begin{aligned}
R(\boldsymbol{\theta}, T) &= R_B(T) + R_C(T) \\
&\leq \sum_{m \neq m^*} \left(\frac{\alpha_m \log(T)}{\psi_m(\Delta_m/2)} + \frac{2}{\alpha - 2} \right) \\
&\quad + \frac{2}{\exp\left(\frac{2}{K_{m^*}} \left(\frac{\delta}{\bar{D}_{1,m^*}} \right)^{2\gamma_{1,m^*}} \right) - 1} \\
&= O(\log(T)),
\end{aligned}$$

which completes the proof. □

2 Proof of Theorem 6

First we analyze the worse-case regret caused by choosing the suboptimal groups and it has been derived that $\mathbb{E}[N_m(T)] \leq \frac{\alpha_m \log T}{\psi(\Delta_m/2)} = C_m \frac{\log T}{\Delta_m^{\xi_m}}$. With $\xi = \max_{m \in \mathcal{M}} \xi_m$, we now rewrite the regret without dependence on the sub-optimality gap as:

$$\begin{aligned}
R_B^i(T) &= \sum_{m:\Delta_m > 0} \Delta_m \mathbb{E}[N_m(T)] \\
&= \sum_{m:\Delta_m < \Delta} \Delta_m \mathbb{E}[N_m(T)] + \sum_{m:\Delta_m \geq \Delta} \Delta_m \mathbb{E}[N_m(T)] \\
&< T\Delta + CM \frac{\log T}{\Delta^{\xi}} \\
&= C_1 (M \log T)^{\xi} T^{1-\xi},
\end{aligned}$$

where in the last step Δ is chosen to be $(CM \log T / T)^{\xi}$ to optimize the upper bound.

Next, we analyze the regret caused by the suboptimal arms within the optimal group. From inequality (10), we have:

$$\begin{aligned}
&\mathbb{E}[|\hat{X}_{m^*,\hat{k}_{m^*}}(t) - X_{m^*,\hat{k}_{m^*}}|] = \\
&\int_0^1 \mathbb{P}(|\hat{X}_{m^*,\hat{k}_{m^*}}(t) - X_{m^*,\hat{k}_{m^*}}| > x) \mathrm{d}x \\
&= \int_0^{\infty} 2\exp(-2x^2 N_{m^*,\hat{k}_{m^*}}) \mathrm{d}x \\
&\leq \sqrt{\frac{\pi K_{m^*}}{2N_{m^*}(t)}}.
\end{aligned}$$

From Assumption 1, Proposition 2 and inequality (6), the following can be derived:

$$\begin{aligned}
&|\mu_{m^*}(\hat{\theta}^t_{m^*}) - \mu_{m^*}(\theta^t_{m^*})| \\
&\leq D_{2,m^*} \bar{D}_{1,m^*}^{\gamma_{2,m^*}} |\hat{X}_{m^*,\hat{k}_{m^*}}(t) - \mu_{m^*,\hat{k}_{m^*}}(\theta^t_{m^*})|^{\bar{\gamma}_{1,m^*} \gamma_{2,m^*}} \\
&= \phi(|\hat{X}_{m^*,\hat{k}_{m^*}}(t) - \mu_{m^*,\hat{k}_{m^*}}(\theta^t_{m^*})|) \qquad (12) \\
&< \phi\left(\sqrt{\frac{\pi K_{m^*}}{2N_{m^*}(t)}} \right). \qquad (13)
\end{aligned}$$

Adding up to time budget T, we get:

$$\begin{aligned}
R_C^i(T) &\leq \sum_{t=1}^{T} \phi\left(\sqrt{\frac{\pi K_{m^*}}{2t}} \right) \\
&\leq \frac{D_{2,m^*} \bar{D}_{1,m^*}^{\gamma_{2,m^*}} (\pi K_{m^*})^{\xi_{m^*}}}{1 - \xi_{m^*}} T^{\xi_{m^*}}.
\end{aligned}$$

Combining the bounds for $R_B^i(T)$ and $R_C^i(T)$ completes the proof.

3 Proof of Lemma 10

We first consider the probability of choosing arm 2 in a single time step t under a full information setting.

$$\begin{aligned}\mathbb{P}(\hat{\theta}_t \in \Theta_2^*) &= \mathbb{P}(\hat{\theta}_t > \theta_*)\\ &= \frac{1}{2}(\mathbb{P}(\mu_1^{-1}(\hat{X}_1(t)) > \theta_*)) + \mathbb{P}(\mu_2^{-1}(\hat{X}_2(t)) > \theta_*))\\ &= \frac{1}{2}(\mathbb{P}(\hat{X}_1(t) < \mu_1(\theta_*)) + \mathbb{P}(\hat{X}_2(t) > \mu_2(\theta_*)). \quad (14)\end{aligned}$$

We define two joint distributions as $\nu_1 = \mathcal{N}(\mu_1(\theta), 1) \otimes \mathcal{N}(\mu_1(\theta_*), 1)$ and $\nu_2 = \mathcal{N}(\mu_2(\theta), 1) \otimes \mathcal{N}(\mu_2(\theta_*), 1)$. Inequality (14) therefore can be re-written as:

$$\begin{aligned}\mathbb{P}(\hat{\theta}_t \in \Theta_2^*) &= \frac{1}{2}(\mathbb{P}_{\nu_1}(I_t = 2) + \mathbb{P}_{\nu_2}(I_t = 1))\\ &\geq \frac{1}{4}\exp(-K(\nu_1^{\otimes t}, \nu_2^{\otimes t}))\\ &= \frac{1}{4}\exp\left(-tK(\mu_1(\theta), \mu_2(\theta))\right). \quad (15)\end{aligned}$$

Adding up to time budget T, we have:

$$\begin{aligned}\mathbb{E}(N_2(T)) &\geq \frac{1}{4}\sum_{t=1}^{T}\exp\left(-tK(\mu_1(\theta), \mu_2(\theta))\right)\\ &\geq \frac{1}{4K(\mu_1(\theta), \mu_2(\theta))}(1 - e^{-(T-1)K(\mu_1(\theta), \mu_2(\theta))^2}),\end{aligned}$$

and

$$\lim_{T\to\infty}\mathbb{E}(N_2(T)) \geq \frac{1}{4K(\mu_1(\theta), \mu_2(\theta))},$$

which concludes the proof.

4 Proof of Theorem 7

To get a lower bound of the total regret, we can separately bound the aforementioned two types of suboptimal arms. First, regret due to selecting a suboptimal arm in the optimal group can be bounded as $\Omega(T^{1-\xi})$. From the proof of Lemma 10, inequality (15) can be derived in the two-armed setting. We specify the reward functions as $\mu_1(\theta) = \theta^\gamma$ and $\mu_2(\theta) = 1 - \theta^\gamma$, where $\gamma \geq 1$. When $\theta = \theta_* = 2^{-\frac{1}{\gamma}}$ the two functions lead to the same performance. Define $\theta = \theta_* + \Delta$, and further define the two processes as: $\nu_1 = \mathcal{N}((\theta_* + \Delta)^\gamma, 1) \otimes \mathcal{N}(1/2, 1)$ and $\nu_2 = \mathcal{N}(1 - (\theta_* + \Delta)^\gamma, 1) \otimes \mathcal{N}(1/2, 1)$. We first bound the one step loss as:

$$\begin{aligned}&(\theta_* + \Delta)^\gamma - (1 - (\theta_* + \Delta)^\gamma)\\ =&2(\theta_* + \Delta)^\gamma - 1\\ =&2\theta_*^\gamma(1 + \frac{\Delta}{\gamma})^\gamma - 1\\ =&2\gamma 2^{\frac{1-\gamma}{\gamma}}\Delta + o(\Delta)\\ \geq&b\Delta.\end{aligned}$$

Therefore the cumulative regret is:

$$\begin{aligned}R_B(T) &\geq \frac{b\Delta}{4}\sum_{t=1}^{T}\exp\left(-\frac{t(2(\theta_* + \Delta)^\gamma - 1)^2}{2}\right)\\ &\geq \frac{b\Delta}{4}\sum_{t=1}^{T}\exp(-t2(\theta_* + \Delta)^2\gamma) \quad (16)\\ &\geq b\frac{1 - e^{-16}}{8}\frac{\Delta}{(\Delta + 2^{-\frac{1}{\gamma}})^{2\gamma}},\end{aligned}$$

where inequality (16) is obtained using $(x - y)^\gamma \leq x^\gamma - y^\gamma$ with $\gamma \geq 1$. If we set $\Delta = 2^{-\frac{1}{\gamma}}T^{-\frac{1}{2\gamma}}$, which can be relatively small when T is large, we have $\gamma_1 = \gamma$ and $\gamma_2 = 1$, and the following lower bound:

$$R_B(T) = \Omega(T^{1-\frac{1}{2\gamma}}) = \Omega(T^{1-\xi}). \quad (17)$$

For suboptimal arms in suboptimal groups, the lower bound follows the result in [5], which indicates that:

$$R_C(T) = \Omega(\sqrt{MT}). \quad (18)$$

Combining the two bounds (17) and (18) completes the proof.

5 Proof of Theorem 14

The proof follows the general analysis in Section 4.2 of the main paper, incorporating the window size τ_w.

I. Analyze the regret caused by choosing a suboptimal group We have:

$$\mu_{m^*}(\hat{\theta}_{m^*}(t)) + c_{m^*}(t, \tau_w) \leq \mu_{m^*}(\theta_{m^*}^t), \quad (19)$$

$$\mu_m(\hat{\theta}_m(t)) \geq c_m(t, \tau_w) + \mu_m(\theta_m^t), \quad (20)$$

$$N_m(t-1) < B(\tau_w, \Delta_m),$$

$$B(\tau_w, \Delta_m) = K\log(t \wedge \tau_w)(\frac{2D_{2,m}\bar{D}_{1,m}^{\gamma_{2,m}}}{\Delta_m})^{\frac{2}{\gamma_{1,m}\gamma_{2,m}}}.$$

Now, to show that performances of all groups can be accurately estimated with high probability, we consider inequality (13). Note that $|\hat{X}_{m,k}(t) - \mu_{m,k}(\theta_m^t)|$ is separated as the instantaneous fluctuation of the random variable $X_{m,k}(s)$ around $\mu_{m,k}(\theta_m^s)$ and the bias caused by drifting of parameter $\mu_{m,k}(\theta_m^s) - \mu_{m,k}(\theta_m^t)$ at time s, which can be written as:

$$\begin{aligned}&|\hat{X}_{m,k}(t) - \mu_{m,k}(\theta_m^t)|\\ \leq&\left|\frac{\sum_{\substack{s=t-\tau_w+1\\ m(s)=m\\ k(s)=k}}^{t}(X_{m,k}(s) - \mu_{m,k}(\theta_m^s))}{N_{m,k}(t, \tau_w)}\right| +\end{aligned}$$

$$\frac{\sum_{\substack{s=t-\tau_w+1\\ m(s)=m\\ k(s)=k}}^{t} |\mu_{m,k}(\theta_m^s) - \mu_{m,k}(\theta_m^t)|}{N_{m,k}(t,\tau_w)} \tag{21}$$

$$\leq \left| \frac{\sum_{\substack{s=t-\tau_w+1\\ m(s)=m\\ k(s)=k}}^{t} (X_{m,k}(s) - \mu_{m,k}(\theta_m^s))}{N_{m,k}(t,\tau_w)} \right| + \frac{\sum_{\substack{s=t-\tau_w+1\\ m(s)=m\\ k(s)=k}}^{t} 2D_{2,m}|\theta_m^s - \theta_m^t|^{\gamma_{2,m}}}{N_{m,k}(t,\tau_w)}. \tag{22}$$

The bias component in the second inequality (22) is based on Proposition 2. With Assumption 12 and the fact that $|t-s| \leq \tau_w$, the bias is bounded as a function of τ_w. If we adjust the length of the window, the bias can be controlled to be smaller than half of the projected padding function, leaving the other half to be used in bounding the fluctuations. This is formally written as:

$$\begin{aligned}
&\mathbb{P}_{\tau_w}(\mu_m(\hat{\theta}_m(t)) > \mu_m(\theta_m^t) + c_m(t,\tau_w)) \\
&\leq \mathbb{P}_{\tau_w}\left(|\hat{X}_{m,k}(t) - \mu_{m,k}(\theta_m^t)| > \phi_m^{-1}(c_m(t,\tau_w)) \right) \\
&\leq \mathbb{P}_{\tau_w}\left(\left| \frac{\sum_{\substack{s=t-\tau_w+1\\ k(s)=k}}^{t} (X_{m,k}(s) - \mu_{m,k}(\theta_m^s))}{N_{m,k}(t,\tau_w)} \right| > \frac{1}{2}\phi_m^{-1}(c_m(t,\tau_w)) \right) \\
&\leq \exp\left(-2\frac{\alpha_m \log(t \wedge \tau_w)}{N_m(t,\tau_w)} \frac{N_m(t,\tau_w)}{K_m} \right) \\
&\leq \left(\frac{1}{t \wedge \tau_w} \right)^{\frac{2K}{K_m}}.
\end{aligned}$$

II. Analyze the regret by choosing a suboptimal arm in the optimal group

Consider the minmax bound for the fluctuations, we have [1]:

$$\begin{aligned}
&\mathbb{E}_{\tau_w}\left[\left| \frac{\sum_{\substack{s=t-\tau_w+1\\ k(s)=k}}^{t} (X_{m,k}(s) - \mu_{m,k}(\theta_m^s))}{N_{m,k}(t,\tau_w)} \right| \right] \\
&= \int_0^\infty \mathbb{P}\left(\left| \frac{\sum_{\substack{s=t-\tau_w+1\\ k(s)=k}}^{t} (X_{m,k}(s) - \mu_{m,k}(\theta_m^s))}{N_{m,k}(t,\tau_w)} \right| > x \right) \mathrm{d}x \\
&\leq \int_0^\infty 2\exp(-x^2 N_{m,k}(t,\tau_w))\mathrm{d}x \\
&= \sqrt{\frac{\pi}{N_{m,k}(t,\tau_w)}}.
\end{aligned}$$

Therefore, the regret from choosing a suboptimal arm within the optimal group is bounded as:

$$\begin{aligned}
\mathbb{E}[r_t(\theta_{m^*}^t)] &\leq \phi(|\hat{X}_{m^*,\hat{k}_{m^*}}(t) - \mu_{m^*,\hat{k}_{m^*}}(\theta_{m^*}^t)|) \\
&\leq \phi_{m^*}(\sqrt{\frac{\pi K_{m^*}}{\tau_w}} + 2D_{2,m^*}\left(\frac{\tau_w}{\tau}\right)^{\gamma_2}).
\end{aligned}$$

Combining the above with the confusing period, and denote $X = \max D_{2,m}\bar{D}_{1,m}^{\gamma_{2,m}}, \gamma_2 = \max \gamma_{2,m}, \bar{\gamma}_1 = \max \bar{\gamma}_1, \alpha_m = \min \frac{2K}{K_m} = 2, \Delta = \min \Delta_m$, we have:

$$\begin{aligned}
&R(T) \leq \Delta G(\Delta,T)M + M\lceil T/\tau_w \rceil B(\tau_w) + \\
&2\sum_{m=1}^{M}\sum_{t=1}^{T}\left(\frac{1}{t \wedge \tau_w}\right)^{\alpha} + T\phi_{m^*}(\sqrt{\frac{\pi K_{m^*}}{\tau_w}} + 2D_2\left(\frac{\tau_w}{\tau}\right)^{\gamma_2}) \\
&\leq M + M\lceil T/\tau_w \rceil (\frac{2X}{\Delta})^{\frac{1}{\xi}} \log(\tau_w) + MT(1-\tau_w^{-1}) + \\
&2MT\tau_w^{-2} + TX(\pi K_{m^*})^{\xi}\tau_w^{-\xi} + TX(2D_2)^{\bar{\gamma}_1\gamma_2}\left(\frac{\tau_w}{\tau}\right)^{\bar{\gamma}_1\gamma_2^2}.
\end{aligned}$$

Set $\tau_w = \max \tau^{\frac{2\gamma_{2,m}}{2\gamma_{2,m}+1}}$, which satisfies the constraint on the bias, the regret per unit time of SW-UCB-g satisfies:

$$\begin{aligned}
\lim_{T\to\infty}\frac{R(T)}{T} &\leq C_1\tau^{-\frac{\bar{\gamma}_1\gamma_2^2}{2\gamma_2+1}} + C_2\tau^{-\frac{2\gamma_2}{2\gamma_2+1}}\log(\tau) + \\
&2M\tau^{-\frac{4\gamma_2^2}{2\gamma_2+1}} + 2M(1-\alpha_m^{-\frac{2\gamma_2}{2\gamma_2+1}}) + f(M).
\end{aligned} \tag{23}$$

The proof is complete by taking the scaling order of (23).