Supplement to "Adaptive Exploration in Linear Contextual Bandit"

In Section A, we provide main proofs for asymptotic lower bound and upper bound. In Section B, we prove several main lemmas. In Section C, some supporting lemmas are presented for the sake of completeness.

A Proofs of Asymptotic Lower and Upper Bounds

First of all, we define the sub-optimal action set as $\mathcal{A}_-^m = \mathcal{A}^m \setminus \{x : \Delta_x^m = 0\}$ and denote $\mathcal{A} = \cup_{m=1}^M \mathcal{A}^m$ and $\mathcal{A}_- = \cup_{m=1}^M \mathcal{A}_-^m$.

A.1 Proof of Lemma 3.2

The proof idea follows if $\bar{G}_n$ is not sufficiently large in every direction, then some alternative parameters are not sufficiently identifiable.

Step One. We fix a consistent policy π and fix a context $m \in [M]$ as well as a sub-optimal arm $x \in \mathcal{A}_-^m$. Consider another parameter $\widetilde{\theta} \in \mathbb{R}^d$ such that it is close to θ but x_m^* is not the optimal arm in bandit $\widetilde{\theta}$ for action set $\mathcal{A}^m$. Specifically, we construct

$$\widetilde{\theta} = \theta + \frac{H(x - x_m^*)}{\|x - x_m^*\|_H^2}(\Delta_x^m + \varepsilon),$$

where $H \in \mathbb{R}^{d \times d}$ is some positive semi-definite matrix and $\varepsilon > 0$ is some absolute constant that will be specified later. Since the sub-optimality gap $\widetilde{\Delta}_{x_m^*}^m$ satisfies

$$\langle x - x_m^*, \widetilde{\theta} \rangle = \langle x - x_m^*, \theta \rangle + \Delta_x^m + \varepsilon = \varepsilon > 0, \tag{A.1}$$

it ensures that x_m^* is ε-suboptimal in bandit $\widetilde{\theta}$.

We define $T_x(n) = \sum_{t=1}^n \mathbb{I}(X_t = x)$ and let $\mathbb{P}$ and $\widetilde{\mathbb{P}}$ be the measures on the sequence of outcomes $(X_1, Y_1, \ldots, X_n, Y_n)$ induced by the interaction between the policy and the bandit θ and $\widetilde{\theta}$ respectively. By the definition of $\bar{G}_n$ in (3.2), we have

$$\begin{aligned}
\frac{1}{2}\|\theta - \widetilde{\theta}\|_{\bar{G}_n}^2 &= \frac{1}{2}(\theta - \widetilde{\theta})^\top \bar{G}_n (\theta - \widetilde{\theta}) \\
&= \frac{1}{2}(\theta - \widetilde{\theta})^\top \mathbb{E}\Big[\sum_{x \in \mathcal{A}} T_x(n) x x^\top\Big](\theta - \widetilde{\theta}) \\
&= \frac{1}{2}\sum_{x \in \mathcal{A}} \mathbb{E}\Big[T_x(n)\Big]\langle x, \theta - \widetilde{\theta}\rangle^2.
\end{aligned}$$

Applying the Bretagnolle-Huber inequality inequality in Lemma C.1 and divergence decomposition lemma in Lemma C.2, it holds that for any event $\mathcal{D}$,

$$\frac{1}{2}\|\theta - \widetilde{\theta}\|_{\bar{G}_n}^2 = \mathrm{KL}(\mathbb{P}, \widetilde{\mathbb{P}}) \geq \log\Big(\frac{1}{2(\mathbb{P}(\mathcal{D}) + \widetilde{\mathbb{P}}(\mathcal{D}^c))}\Big). \tag{A.2}$$

Step Two. In the following, we start to derive a lower bound of $R_\theta^\pi(n)$,

$$\begin{aligned}
R_\theta^\pi(n) &= \mathbb{E}\Big[\sum_{t=1}^n \langle x_{c_t}^* - X_t, \theta\rangle\Big] = \mathbb{E}\Big[\sum_{m=1}^M \sum_{t:c_t=m} \langle x_m^* - X_t, \theta\rangle\Big] \\
&\geq \mathbb{E}\Big[\sum_{t:c_t=m} \langle x_m^* - X_t, \theta\rangle\Big] = \mathbb{E}\Big[\sum_{t:c_t=m} \Delta_{X_t}^m\Big] \\
&\geq \Delta_{\min}\mathbb{E}\Big[\sum_{t:c_t=m} \mathbb{I}(X_t \neq x_m^*)\Big] = \Delta_{\min}\mathbb{E}\Big[\sum_{t=1}^n \mathbb{I}(c_t = m) - \sum_{t=1}^n \mathbb{I}(c_t = m)\mathbb{I}(X_t = x_m^*)\Big],
\end{aligned}$$

where the first inequality comes from the fact that $\langle x_m^* - X_t, \theta \rangle \geq 0$ for all $m \in [M]$. Define the event $\mathcal{D}$ as follows,

$$\mathcal{D} = \Big\{ \sum_{t=1}^n \mathbb{I}(c_t = m)\mathbb{I}(X_t = x_m^*) \leq \frac{1}{2}\sum_{t=1}^n \mathbb{I}(c_t = m) \Big\}. \tag{A.3}$$

When event $\mathcal{D}$ holds, we will only pull at most half of total rounds for the optimal action of action set m. Then it holds that

$$\begin{aligned} R_\theta^\pi(n) &\geq \Delta_{\min}\mathbb{E}\Big[\Big(\sum_{t=1}^n \mathbb{I}(c_t = m) - \sum_{t=1}^n \mathbb{I}(c_t = m)\mathbb{I}(X_t = x_m^*)\Big)\mathbb{I}(\mathcal{D})\Big] \\ &\geq \Delta_{\min}\mathbb{E}\Big[\frac{1}{2}\sum_{t=1}^n \mathbb{I}(c_t = m)\mathbb{I}(\mathcal{D})\Big]. \end{aligned}$$

Define another event $\mathcal{B}$ as follows,

$$\mathcal{B} = \Big\{\frac{1}{2}\sum_{t=1}^n \mathbb{I}(c_t = m) \geq \frac{np_m}{2} - \delta/2\Big\}, \tag{A.4}$$

where $\delta > 0$ will be chosen later and p_m is the probability that the environment picks context m. From the definition of c_t, we have $\mathbb{E}[\sum_{t=1}^n \mathbb{I}(c_t = m)] = np_m$. By the standard Hoeffding's inequality (Vershynin, 2010), it holds that

$$\mathbb{P}\Big(\frac{1}{2}\sum_{t=1}^n \mathbb{I}(c_t = m) - \frac{np_m}{2} \geq -\frac{\delta}{2}\Big) \geq 1 - \exp(-\frac{2\delta^2}{n}),$$

which implies

$$\mathbb{P}(\mathcal{B}^c) \leq \exp(-2\delta^2/n).$$

By the definition of events $\mathcal{D}, \mathcal{B}$ in (A.3),(A.4), we have

$$\begin{aligned} R_\theta^\pi(n) &\geq \Delta_{\min}\mathbb{E}\Big[\frac{1}{2}\sum_{t=1}^n \mathbb{I}(c_t = m)\mathbb{I}(\mathcal{D})\mathbb{I}(\mathcal{B})\Big] \\ &\geq \Delta_{\min}\mathbb{E}\Big[(\frac{1}{2}np_m - \frac{\delta}{2})\mathbb{I}(\mathcal{D})\mathbb{I}(\mathcal{B})\Big] \\ &= \Delta_{\min}(\frac{1}{2}np_m - \frac{\delta}{2})\mathbb{P}(\mathcal{D} \cap \mathcal{B}) \\ &\geq \Delta_{\min}(\frac{1}{2}np_m - \frac{\delta}{2})(\mathbb{P}(\mathcal{D}) - \mathbb{P}(\mathcal{B}^c)). \end{aligned}$$

Letting $\delta = np_m/2$, we have

$$R_\theta^\pi(n) \geq \Delta_{\min}\frac{np_m}{4}\Big(\mathbb{P}(\mathcal{D}) - \exp(-\frac{np_m^2}{2})\Big). \tag{A.5}$$

On the other hand, we let $\widetilde{\mathbb{E}}$ is taken with respect to probability measures $\widetilde{\mathbb{P}}$. Then $R_{\widetilde{\theta}}^\pi(n)$ can be lower bounded as follows,

$$\begin{aligned} R_{\widetilde{\theta}}^\pi(n) &= \widetilde{\mathbb{E}}\Big[\sum_{m=1}^M\sum_{t=1}^n \mathbb{I}(c_t = m)\widetilde{\Delta}_{X_t}^m\Big] \\ &\geq \widetilde{\mathbb{E}}\Big[\sum_{t=1}^n \mathbb{I}(c_t = m)\mathbb{I}(X_t = x_m^*)\Big]\widetilde{\Delta}_{x_m^*}^m, \end{aligned}$$

where we throw out all the sub-optimality gap terms except $\widetilde{\Delta}^m_{x^*_m}$. Using the fact that $\widetilde{\Delta}^m_{x^*_m}$ is ε-suboptimal, it holds that

$$\begin{aligned}
R^\pi_{\widetilde{\theta}}(n) &\geq \varepsilon\widetilde{\mathbb{E}}\Big[(\sum_{t=1}^n \mathbb{I}(c_t = m)\mathbb{I}(X_t = x^*_m))\mathbb{I}(\mathcal{D}^c)\Big] \\
&> \varepsilon\widetilde{\mathbb{E}}\Big[\frac{1}{2}\sum_{t=1}^n \mathbb{I}(c_t = m)\mathbb{I}(\mathcal{D}^c)\Big] \\
&\geq \varepsilon\widetilde{\mathbb{E}}\Big[\frac{1}{2}\sum_{t=1}^n \mathbb{I}(c_t = m)\mathbb{I}(\mathcal{D}^c)\mathbb{I}(\mathcal{B})\Big] \\
&\geq \varepsilon(\frac{np_m}{2} - \frac{\delta}{2})\widetilde{\mathbb{P}}(\mathcal{D}^c \cap \mathcal{B}) \\
&\geq \varepsilon(\frac{np_m}{2} - \frac{\delta}{2})(\widetilde{\mathbb{P}}(\mathcal{D}^c) - \widetilde{\mathbb{P}}(\mathcal{B}^c)) \\
&\geq \varepsilon(\frac{np_m}{2} - \frac{\delta}{2})(\widetilde{\mathbb{P}}(\mathcal{D}^c) - \exp(-\frac{2\delta^2}{n})) \\
&= \varepsilon\frac{np_m}{4}\widetilde{\mathbb{P}}(\mathcal{D}^c) - \varepsilon\frac{np_m}{4}\exp(-\frac{np_m^2}{2}). \qquad \text{(A.6)}
\end{aligned}$$

Now we have derived the lower bounds (A.5)(A.6) for $R^\pi_\theta(n), R^\pi_{\widetilde{\theta}}(n)$ respectively.

Step Three. Combining the lower bounds of $R^\pi_\theta(n)$ and $R^\pi_{\widetilde{\theta}}(n)$ together, it holds that

$$R^\pi_\theta(n) + R^\pi_{\widetilde{\theta}}(n) \geq \frac{np_m}{4}\Big(\mathbb{P}(\mathcal{D})\Delta_{\min} + \widetilde{\mathbb{P}}(\mathcal{D}^c)\varepsilon\Big) - \frac{np_m}{4}\exp(-\frac{np_m^2}{2})(\varepsilon + \Delta_{\min}).$$

Letting $\varepsilon \leq \Delta_{\min}$, we have

$$R^\pi_\theta(n) + R^\pi_{\widetilde{\theta}}(n) \geq \varepsilon\frac{np_m}{4}\Big(\mathbb{P}(\mathcal{D}) + \widetilde{\mathbb{P}}(\mathcal{D}^c)\Big) - \frac{np_m}{4}\exp(-\frac{np_m^2}{2})2\Delta_{\min}.$$

This implies

$$\frac{R^\pi_\theta(n) + R^\pi_{\widetilde{\theta}}(n)}{\varepsilon np_m/4} + \frac{1}{\varepsilon}\exp(-\frac{np_m^2}{2})2\Delta_{\min} \geq \mathbb{P}(\mathcal{D}) + \widetilde{\mathbb{P}}(\mathcal{D}^c). \qquad \text{(A.7)}$$

Plugging (A.7) into (A.2), we have

$$\begin{aligned}
\frac{1}{2}\|\theta - \widetilde{\theta}\|^2_{\bar{G}_n} &\geq \log\Big(\frac{1}{2(\mathbb{P}(\mathcal{D}) + \widetilde{\mathbb{P}}(\mathcal{D}^c))}\Big) \\
&\geq \log\Big(\frac{1}{\frac{R^\pi_\theta(n)+R^\pi_{\widetilde{\theta}}(n)}{\varepsilon np_m/8} + \frac{1}{\varepsilon}\exp(-\frac{np_m^2}{2})4\Delta_{\min}}\Big) \\
&= \log\Big(\frac{n}{\frac{R^\pi_\theta(n)+R^\pi_{\widetilde{\theta}}(n)}{\varepsilon p_m/8} + \frac{n}{\varepsilon}\exp(-\frac{np_m^2}{2})4\Delta_{\min}}\Big) \\
&= \log(n) - \log\Big(\frac{R^\pi_\theta(n) + R^\pi_{\widetilde{\theta}}(n)}{\varepsilon p_m/8} + \frac{4n}{\varepsilon}\exp(-\frac{np_m^2}{2})\Delta_{\min}\Big).
\end{aligned}$$

Dividing by $\log(n)$ for both sides, we reach

$$\frac{\|\theta - \widetilde{\theta}\|^2_{\bar{G}_n}}{2\log(n)} \geq 1 - \frac{\log\Big(\frac{R^\pi_\theta(n)+R^\pi_{\widetilde{\theta}}(n)}{\varepsilon p_m/8} + \frac{4n}{\varepsilon}\exp(-\frac{np_m^2}{2})\Delta_{\min}\Big)}{\log(n)}.$$

From the definition of consistent policies (3.1), it holds that

$$\limsup_{n\to\infty}\frac{\log(R^\pi_\theta(n) + R^\pi_{\widetilde{\theta}}(n))}{\log(n)} \leq 0.$$

In addition, by using the fact that $\lim_{n\to\infty} n\exp(-n) = 0$, it follows that

$$\liminf_{n\to\infty} \frac{\|\theta - \widetilde{\theta}\|^2_{\bar{G}_n}}{2\log(n)} \geq 1. \tag{A.8}$$

Step Four. Let's denote

$$\rho_n(H) = \frac{\|x - x^*_m\|^2_{\bar{G}_n^{-1}} \|x - x^*_m\|^2_{H\bar{G}_n H}}{\|x - x^*_m\|^4_H}.$$

Then we can rewrite

$$\frac{1}{2}\|\theta - \widetilde{\theta}\|^2_{\bar{G}_n} = \frac{(\Delta^m_x + \varepsilon)^2}{2\|x - x^*_m\|^2_{\bar{G}_n^{-1}}} \rho_n(H).$$

Plugging this into (A.8) and letting ε to zero, we see that

$$\liminf_{n\to\infty} \frac{\rho_n(H)}{\|x - x^*_m\|^2_{\bar{G}_n^{-1}} \log(n)} \geq \frac{2}{(\Delta^m_x)^2}. \tag{A.9}$$

Now, we consider the following lemma, extracted from the proof of Theorem 25.1 of the book by Lattimore and Szepesvári (2019). The detailed proof is deferred to Section B.6.

Lemma A.1. Let $\{G_n\}_{n\geq 0}$ be a sequence of $d\times d$ positive definite matrices, $s \in \mathbb{R}^d$. For H positive semi-definite $d\times d$ matrix such that $\|s\|_H > 0$ and $n \geq 0$, let $\rho_n(H) = \frac{\|s\|^2_{G_n^{-1}}\|s\|^2_{HG_nH}}{\|s\|^4_H}$. Assume that $\liminf_{n\to\infty} \frac{\lambda_{\min}(G_n)}{\log(n)} > 0$ and that for some $c > 0$,

$$\liminf_{n\to\infty} \frac{\rho_n(H)}{\|s\|^2_{G_n^{-1}} \log(n)} \geq c. \tag{A.10}$$

Then, $\limsup_{n\to\infty} \log(n)\|s\|^2_{G_n^{-1}} \leq 1/c$.

The proof of $\liminf_{n\to\infty} \frac{\lambda_{\min}(G_n)}{\log(n)} > 0$ could refer Appendix C in Lattimore and Szepesvári (2017). Clearly, this lemma with $G_n = \bar{G}_n$, $c = 2/(\Delta^m_x)^2$, $H = \lim_{n\to\infty} \bar{G}_n^{-1}/\|\bar{G}_n^{-1}\|$ and $s = x - x^*_m$ gives the desired statement.

■

A.2 Proof of Theorem 4.3: Asymptotic Upper Bound

We write $\Delta_{\max} = \max_{x,m} \Delta^m_x$ and abbreviate $R(n) = R^\pi_\theta(n)$. From the design of the initialisation, G_t is guaranteed to be invertible since each $\mathcal{A}^m$ is assumed to span $\mathbb{R}^d$. The regret during the initialisation is at most $d\Delta_{\max} \approx o(\log(n))$ and thus we ignore the regret during initialisation in the following.

First, we introduce a refined concentration inequality for the least square estimator constructed by adaptive data. The proof could refer to the proof of Theorem 8 in Lattimore and Szepesvári (2017).

Lemma A.2. Suppose for $t \geq d$, G_t is invertible. For any $\delta \in (0,1)$, we have

$$\mathbb{P}\Big(\exists t \geq d, \exists x \in \mathcal{A}, \text{such that } \big|\langle x, \widehat{\theta}_t\rangle - \langle x, \theta\rangle\big| \geq \|x\|_{G_t^{-1}} f^{1/2}_{n,\delta}\Big) \leq \delta,$$

and

$$f_{n,\delta} = 2\Big(1 + \frac{1}{\log(n)}\Big)\log(1/\delta) + cd\log(d\log(n)), \tag{A.11}$$

where $c > 0$ is some universal constant. We write $f_n = f_{n,1/n}$ for short.

Let us define the event $\mathcal{B}_t$ as follows

$$\mathcal{B}_t = \Big\{\exists t \geq d, \exists x \in \mathcal{A}, \text{such that } |x^\top \widehat{\theta}_t - x^\top \theta| \geq \|x\|_{G_t^{-1}} f_n^{1/2}\Big\}. \tag{A.12}$$

From Lemma A.2, we have $\mathbb{P}(\mathcal{B}_t) \leq 1/n$ by choosing $\delta = 1/n$. We decompose the cumulative regret with respect to event $\mathcal{B}_t$ as follows,

$$\begin{aligned} R(n) &= \mathbb{E}\Big[\sum_{t=1}^{n} \sum_{x \in \mathcal{A}_-^{c_t}} \Delta_x^{c_t} \mathbb{I}(X_t = x)\Big] \\ &= \mathbb{E}\Big[\sum_{t=1}^{n} \sum_{x \in \mathcal{A}_-^{c_t}} \Delta_x^{c_t} \mathbb{I}(X_t = x, \mathcal{B}_t)\Big] + \mathbb{E}\Big[\sum_{t=1}^{n} \sum_{x \in \mathcal{A}_-^{c_t}} \Delta_x^{c_t} \mathbb{I}(X_t = x, \mathcal{B}_t^c)\Big]. \end{aligned} \tag{A.13}$$

To bound the first term in (A.13), we observe that

$$\begin{aligned} &\limsup_{n\to\infty} \frac{\mathbb{E}\Big[\sum_{t=1}^{n} \sum_{x \in \mathcal{A}_-^{c_t}} \Delta_x^{c_t} \mathbb{I}(X_t = x, \mathcal{B}_t)\Big]}{\log(n)} \\ = &\limsup_{n\to\infty} \frac{\mathbb{E}\Big[\sum_{t=1}^{n} \Delta_{X_t}^{c_t} \mathbb{I}(\mathcal{B}_t)\Big]}{\log(n)} \leq \limsup_{n\to\infty} \frac{\Delta_{\max} \sum_{t=1}^{n} \mathbb{P}(\mathcal{B}_t)}{\log(n)} = \limsup_{n\to\infty} \frac{\Delta_{\max} \sum_{t=1}^{n} \frac{1}{n}}{\log(n)} \\ = &\limsup_{n\to\infty} \frac{\Delta_{\max}}{\log(n)} = 0. \end{aligned} \tag{A.14}$$

To bound the second term in (A.13), we define the event $\mathcal{D}_{t,c_t}$ as follows,

$$\mathcal{D}_{t,c_t} = \left\{\forall x \in \mathcal{A}^{c_t}, \|x\|_{G_t^{-1}}^2 \leq \max\left\{\frac{(\widehat{\Delta}_{\min}(t))^2}{f_n}, \frac{(\Delta_x^{c_t}(t))^2}{f_n}\right\}\right\}. \tag{A.15}$$

When $\mathcal{D}_{t,c_t}$ occurs, the algorithm exploits at round t. Otherwise, the algorithm explores at round t. We decompose the second term in (A.13) as the exploitation regret and exploration regret:

$$\begin{aligned} &\mathbb{E}\Big[\sum_{t=1}^{n} \sum_{x \in \mathcal{A}_-^{c_t}} \Delta_x^{c_t} \mathbb{I}(X_t = x, \mathcal{B}_t^c)\Big] \\ = &\mathbb{E}\Big[\sum_{t=1}^{n} \sum_{x \in \mathcal{A}_-^{c_t}} \Delta_x^{c_t} \mathbb{I}(X_t = x, \mathcal{B}_t^c, \mathcal{D}_{t,c_t})\Big] + \mathbb{E}\Big[\sum_{t=1}^{n} \sum_{x \in \mathcal{A}_-^{c_t}} \Delta_x^{c_t} \mathbb{I}(X_t = x, \mathcal{B}_t^c, \mathcal{D}_{t,c_t}^c)\Big]. \end{aligned} \tag{A.16}$$

We bound those two terms in Lemmas A.3-A.4 respectively.

Lemma A.3. The exploitation regret satisfies

$$\limsup_{n\to\infty} \frac{\mathbb{E}\Big[\sum_{t=1}^{n} \sum_{x \in \mathcal{A}_-^{c_t}} \Delta_x \mathbb{I}(X_t = x, \mathcal{B}_t^c, \mathcal{D}_{t,c_t})\Big]}{\log(n)} = 0 \tag{A.17}$$

Lemma A.4. The exploration regret satisfies

$$\limsup_{n\to\infty} \frac{\mathbb{E}\Big[\sum_{t=1}^{n} \sum_{x \in \mathcal{A}_-^{c_t}} \Delta_x \mathbb{I}(X_t = x, \mathcal{B}_t^c, \mathcal{D}_{t,c_t}^c)\Big]}{\log(n)} \leq \mathcal{C}(\theta, \mathcal{A}^1, \ldots, \mathcal{A}^M), \tag{A.18}$$

where $\mathcal{C}(\theta, \mathcal{A}^1, \ldots, \mathcal{A}^M)$ is defined in Theorem 3.3.

Combining Lemmas A.3-A.4 together, we reach our conclusion. ■

B Proofs of Several lemmas

B.1 Proof of Lemma A.3: Exploitation Regret

When $\mathcal{B}_t^c$ defined in (A.12) occurs, we have

$$\max_{x\in\mathcal{A}} \big|\langle \widehat{\theta}_t - \theta, x\rangle\big| \le \|x\|_{G_t^{-1}} f_n^{1/2}. \tag{B.1}$$

When $\mathcal{D}_{t,m}$ defined in (A.15) occurs, we have

$$\|x\|^2_{G_t^{-1}} \le \max\Big\{\frac{\widehat{\Delta}^2_{\min}(t)}{f_n}, \frac{(\widehat{\Delta}_x^m(t))^2}{f_n}\Big\} = \frac{(\widehat{\Delta}_x^m(t))^2}{f_n}, \tag{B.2}$$

holds for any action $x \in \mathcal{A}^m$ and $\widehat{\Delta}_x^m(t) > 0$. If $x_m^* = \widehat{x}_m^*(t)$, there is no regret occurred. Otherwise, putting (B.1) and (B.2) together with the optimal action x_m^*, it holds that

$$|\langle \widehat{\theta}_t - \theta, x_m^*\rangle| \le \|x_m^*\|_{G_t^{-1}} f_n^{1/2} \le \widehat{\Delta}^m_{x_m^*}(t). \tag{B.3}$$

We decompose the sub-optimality gap of $\widehat{x}_m^*(t)$ as follows,

$$\begin{aligned}
& \langle x_m^*, \theta\rangle - \langle \widehat{x}_m^*(t), \theta\rangle \\
= \;& \langle x_m^*, \theta - \widehat{\theta}_t\rangle + \langle x_m^*, \widehat{\theta}_t\rangle - \langle \widehat{x}_m^*(t), \theta - \widehat{\theta}_t\rangle - \langle \widehat{x}_m^*(t), \widehat{\theta}_t\rangle \\
= \;& \langle x_m^*, \theta - \widehat{\theta}_t\rangle - \widehat{\Delta}^m_{x_m^*}(t) + \langle \widehat{x}_m^*(t), \widehat{\theta}_t - \theta\rangle \\
\le \;& \langle \widehat{x}_m^*(t), \widehat{\theta}_t - \theta\rangle.
\end{aligned} \tag{B.4}$$

For each $x \in \mathcal{A}$, we define

$$\tau_x = \min\Big\{N : \forall t \ge d, \mathcal{D}_{t,c_t} \text{ occurs}, N_x(t) \ge N, \text{implies } |\langle \widehat{\theta}_t - \theta, x\rangle| \le \frac{\Delta_{\min}}{2}\Big\}. \tag{B.5}$$

When $N_{\widehat{x}_m^*(t)}(t) \ge \tau_{\widehat{x}_m^*(t)}$, it holds that

$$|\langle \widehat{\theta}_t - \theta, \widehat{x}_m^*(t)\rangle| \le \frac{\Delta_{\min}}{2}.$$

Together with (B.4), we have

$$\langle x_m^*, \theta\rangle - \langle \widehat{x}_m^*(t), \theta\rangle \le \frac{\Delta_{\min}}{2}.$$

Combining this with the fact that the instantaneous regret either vanishes or is larger than $\Delta_{\min}$, it indicates $x_m^* = \widehat{x}_m^*(t)$. Therefore, we can decompose the exploitation regret with respect to event $\{N_{\widehat{x}_m^*(t)}(t) \ge \tau_{\widehat{x}_m^*(t)}\}$ as follows,

$$\begin{aligned}
& \mathbb{E}\Big[\sum_{t=1}^n \sum_{x\in\mathcal{A}_-^{c_t}} \Delta_x^{c_t} \mathbb{I}(X_t = x, \mathcal{B}_t^c, \mathcal{D}_{t,c_t})\Big] \\
\le \;& \mathbb{E}\Big[\sum_{m=1}^M \sum_{t=1}^n \sum_{x\in\mathcal{A}_-^m} \Delta_x^m \mathbb{I}\Big(X_t = x, \mathcal{B}_t^c, \mathcal{D}_{t,m}, N_{\widehat{x}_m^*(t)}(t) \ge \tau_{\widehat{x}_m^*(t)}\Big)\Big] \\
+ \;& \mathbb{E}\Big[\sum_{m=1}^M \sum_{t=1}^n \sum_{x\in\mathcal{A}_-^m} \Delta_x^m \mathbb{I}\Big(X_t = x, \mathcal{B}_t^c, \mathcal{D}_{t,m}, N_{\widehat{x}_m^*(t)}(t) < \tau_{\widehat{x}_m^*(t)}\Big)\Big].
\end{aligned} \tag{B.6}$$

During exploiting the algorithm always executes the greedy action. When $x_m^* = \widehat{x}_m^*(t)$ the first term in (B.6) results in no regret. For the second term in (B.6), we have

$$
\begin{aligned}
& \mathbb{E}\Big[\sum_{m=1}^{M}\sum_{t=1}^{n}\sum_{x\in\mathcal{A}_-^m} \Delta_x^m \mathbb{I}\Big(X_t = x, \mathcal{B}_t^c, \mathcal{D}_{t,m}, N_{\widehat{x}_m^*(t)} < \tau_{\widehat{x}_m^*(t)}\Big)\Big] \\
\leq \quad & \mathbb{E}\Big[\sum_{m=1}^{M}\sum_{t=1}^{n}\mathbb{I}\Big(\mathcal{B}_t^c, \mathcal{D}_{t,m}, N_{\widehat{x}_m^*(t)}(t) < \tau_{\widehat{x}_m^*(t)}\Big)\Big]\Delta_{\max} \\
\leq \quad & \sum_{m=1}^{M}\sum_{x\in\mathcal{A}}\mathbb{E}(\tau_x)\Delta_{\max} \leq \sum_{x\in\mathcal{A}}\mathbb{E}[\tau_x]\Delta_{\max}.
\end{aligned}
\tag{B.7}
$$

It remains to bound $\mathbb{E}[\tau_x]$ for any $x \in \mathcal{A}$. Let

$$
\Lambda = \min\Big\{\lambda \geq 1 : \forall t \geq d, |\langle \widehat{\theta}_t - \theta, x\rangle| \leq \|x\|_{G_t^{-1}} f_{n,1/\lambda}^{1/2}\Big\}.
$$

From the definition of τ_x in (B.5), we have

$$
\tau_x \leq \max\Big\{N : (f_{n,1/\lambda}/N)^{1/2} \geq \frac{\Delta_{\min}}{2}\Big\},
$$

which implies $\tau_x \leq 4f_{n,1/\Lambda}/\Delta_{\min}^2$. From Lemma A.2, we know that $\mathbb{P}(\Lambda \geq \lambda) \leq 1/\lambda$, which implies $\mathbb{E}[\log\Lambda] \leq 1$. Overall,

$$
\mathbb{E}[\tau_x] \leq \frac{4\mathbb{E}[f_\Lambda]}{\Delta_{\min}^2} \leq \frac{8(1+1/\log(n)) + 4cd\log(d\log(n))}{\Delta_{\min}^2}.
\tag{B.8}
$$

Combining (B.6)-(B.8) together, we reach

$$
\begin{aligned}
& \limsup_{n\to\infty} \frac{\mathbb{E}\Big[\sum_{t=1}^n \sum_{x\in\mathcal{A}_-^{c_t}} \Delta_x \mathbb{I}(x_t = x, \mathcal{B}_t^c, \mathcal{D}_{t,c_t})\Big]}{\log(n)} \\
\leq & \limsup_{n\to\infty} \frac{|\mathcal{A}|\Delta_{\max}\big(8(1+1/\log(n)) + 4cd\log(d\log(n))\big)}{\Delta_{\min}^2\log(n)} = 0.
\end{aligned}
\tag{B.9}
$$

This ends the proof. ■

B.2 Proof of Lemma A.4: Exploration Regret

If all the actions $x \in \mathcal{A}$ satisfy

$$
N_x(t) \geq \min\Big\{f_n/\widehat{\Delta}_{\min}^2(t), T_x(\widehat{\Delta}(t))\Big\},
\tag{B.10}
$$

the following holds using Lemma C.4,

$$
\|x\|_{G_t^{-1}}^2 \leq \max\Big\{\frac{\widehat{\Delta}_{\min}^2(t)}{f_n}, \frac{(\widehat{\Delta}_x^{c_t}(t))^2}{f_n}\Big\}, \text{ for any } x \in \mathcal{A}.
$$

In other words, this implies if there exists an action x such that (B.10) does not hold, e.g. $\mathcal{D}_{t,c_t}^c$ occurs, there must exist an action $x' \in \mathcal{A}$ (x and x' may not be the identical) satisfying

$$
N_{x'}(t) \leq \min\Big\{f_t/\widehat{\Delta}_{\min}^2(t), T_{x'}(\widehat{\Delta}(t))\Big\}.
$$

Based on the criterion in Algorithm 1, we should explore. However, if x' does not belong to $\mathcal{A}^{c_t}$ and all the actions within $\mathcal{A}^{c_t}$ have been explored sufficiently according to the approximation optimal allocation, this exploration is interpreted as "wasted". To alleviate the regret of the wasted exploration, the algorithm acts optimistically as LinUCB.

Let's define a set that records the index of action sets that has not been fully explored until round t,

$$\mathcal{M}_t = \Big\{m : \exists x \in \mathcal{A}^m, N_x(t) \leq \min\{f_n/\widehat{\Delta}^2_{\min}(t), T_x(\widehat{\Delta}(t))\}\Big\}. \tag{B.11}$$

When $\mathcal{D}^c_{t,c_t}$ occurs, it means that $\mathcal{M}_t \neq \emptyset$. If $\mathcal{D}^c_{t,c_t}$ occurs but c_t does not belong to $\mathcal{M}_t$, the algorithm suffers a wasted exploration. We decompose the exploration regret according to the fact if c_t belongs to $\mathcal{M}_t$,

$$\begin{aligned}
& \mathbb{E}\Big[\sum_{t=1}^n \sum_{x\in\mathcal{A}_-^{c_t}} \Delta_x \mathbb{I}(X_t = x, \mathcal{B}_t^c, \mathcal{D}^c_{t,c_t}\Big] \\
= & \underbrace{\mathbb{E}\Big[\sum_{t=1}^n \sum_{x\in\mathcal{A}_-^{c_t}} \Delta_x \mathbb{I}(X_t = x, \mathcal{B}_t^c, \mathcal{D}^c_{t,c_t}, c_t \in \mathcal{M}_t)\Big]}_{R_{\text{ue}}:\text{unwasted exploration}} \\
& \underbrace{+\,\mathbb{E}\Big[\sum_{t=1}^n \sum_{x\in\mathcal{A}_-^{c_t}} \Delta_x \mathbb{I}(X_t = x, \mathcal{B}_t^c, \mathcal{D}^c_{t,c_t}, c_t \notin \mathcal{M}_t)\Big]}_{R_{\text{we}}:\text{wasted exploration}}.
\end{aligned} \tag{B.12}$$

We will bound the unwasted exploration regret and wasted exploration regret in the following two lemmas respectively.

Lemma B.1. The regret during the unwasted explorations satistifies

$$\limsup_{n\to\infty} \frac{R_{\text{ue}}}{\log(n)} \leq \mathcal{C}(\theta, \mathcal{A}_1, \ldots, \mathcal{A}_M). \tag{B.13}$$

The detailed proof is deferred to Section B.3.

Lemma B.2. The regret during the wasted explorations satisfies

$$\limsup_{n\to\infty} \frac{R_{\text{we}}}{\log(n)} = 0. \tag{B.14}$$

The detailed proof is deferred to Section B.5.

Putting (B.12)-(B.14) together, we reach

$$\limsup_{n\to\infty} \frac{\mathbb{E}\Big[\sum_{t=1}^n \sum_{x\in\mathcal{A}_-^{c_t}} \Delta_x^{c_t} \mathbb{I}(X_t = x, \mathcal{B}_t^c, \mathcal{D}^c_{t,c_t})\Big]}{\log(n)} \leq \mathcal{C}(\theta, \mathcal{A}^1, \ldots, \mathcal{A}^M),$$

which ends the proof.

■

B.3 Proof of Lemma B.1: Unwasted Exploration

First, we derive a lower bound for each $N_x(t)$ during the unwasted exploration. Denote $s(t)$ as the number of rounds for unwasted explorations until round t. Indeed, forced exploration can guarantee a lower bound for $N_x(t)$: $\min_{x\in\mathcal{A}} N_x(t) \geq \varepsilon_t s(t)/2$. We prove this by the contradiction argument. Assume this is not true. There may exist $s(t)/2$ rounds $\{t_1, \ldots, t_{s(t)/2}\} \subset \{1, \ldots, t\}$ such that $\min_{x\in\mathcal{A}} N_x(t) \leq \varepsilon_t s(t)$. After $|\mathcal{A}|$

such rounds, we have $\min_x N_x(t)$ is incremented by at least 1 which implies $\min_x N_x(t) \geq s(t)/(2|\mathcal{A}|)$. If $\varepsilon_t \leq 1/|\mathcal{A}|$, it leads to the contradiction. This is satisfied when t is large since $\varepsilon_t = 1/\log(\log t)$.

Second, we set $\beta_n = 1/\log(\log(n))$ and define

$$\zeta = \min\Big\{s : \forall t \geq s, \forall x \in \mathcal{A}, \text{such that } |\langle x, \widehat{\theta}_t\rangle - \langle x, \theta\rangle| \leq \beta_n\Big\}. \tag{B.15}$$

Then we decompose the regret during unwasted explorations with respect to event $\{s(t) \geq \zeta\}$ as follows,

$$\begin{aligned} R_{\text{ue}} &= \mathbb{E}\Big[\sum_{t=1}^n \sum_{x \in \mathcal{A}_-^{c_t}} \Delta_x \mathbb{I}(X_t = x, \mathcal{B}_t^c, \mathcal{D}_{t,c_t}^c, c_t \in \mathcal{M}_t)\Big] \\ &= \underbrace{\mathbb{E}\Big[\sum_{t=1}^n \sum_{x \in \mathcal{A}_-^{c_t}} \Delta_x \mathbb{I}(X_t = x, \mathcal{B}_t^c, \mathcal{D}_{t,c_t}^c, s(t) \geq \zeta, c_t \in \mathcal{M}_t)\Big]}_{I_1} \\ &\quad \underbrace{+\mathbb{E}\Big[\sum_{t=1}^n \sum_{x \in \mathcal{A}_-^{c_t}} \Delta_x \mathbb{I}(X_t = x, \mathcal{B}_t^c, \mathcal{D}_{t,c_t}^c, s(t) < \zeta, c_t \in \mathcal{M}_t)\Big]}_{I_2}. \end{aligned} \tag{B.16}$$

To bound I_2, we have

$$I_2 = \mathbb{E}\Big[\sum_{t=1}^n \Delta_{X_t} \mathbb{I}(\mathcal{B}_t^c, \mathcal{D}_{t,c_t}^c, c_t \in \mathcal{M}_t, s(t) < \zeta)\Big] \leq \Delta_{\max} \mathbb{E}\Big[\sum_{t=1}^n \mathbb{I}(s(t) < \zeta, c_t \in \mathcal{M}_t, \mathcal{D}_{t,c_t}^c)\Big] \leq \Delta_{\max} \mathbb{E}[\zeta].$$

It remains to bound $\mathbb{E}[\zeta]$. Let's define

$$\Lambda = \min\Big\{\lambda : \forall t : \mathcal{D}_{t,c_t}^c, \forall x \in \mathcal{A}, s(t) \geq s, \text{such that } |\langle x, \widehat{\theta}_t\rangle - \langle x, \theta\rangle| \leq \Big(\frac{2}{\varepsilon_t s(t)} f_{n,1/\lambda}\Big)^{1/2}\Big\}.$$

From the definition of ζ in (B.15), we have

$$\zeta \leq \max\Big\{s : \Big(\frac{f_{n,1/\lambda}}{\varepsilon_t s}\Big)^{1/2} \geq \beta_n\Big\},$$

which implies

$$\zeta \leq \frac{2 f_{n,1/\Lambda}}{\varepsilon_t \beta_n^2}. \tag{B.17}$$

In addition, we define

$$\Lambda' = \min\Big\{\lambda : \forall t \geq d, \forall x \in \mathcal{A}, \text{such that } |\langle x, \widehat{\theta}_t\rangle - \langle x, \theta\rangle| \leq \|x\|_{G_t^{-1}} f_{n,1/\lambda}^{1/2}\Big\}.$$

Using the lower bound of $N_x(t)$, it holds that

$$\|x\|_{G_t^{-1}}^2 \leq \frac{1}{N_x(t)} \leq \frac{2}{\varepsilon_t s(t)}.$$

By Lemma A.2, we have

$$\mathbb{P}\Big(\Lambda \geq \frac{1}{\delta}\Big) \leq \mathbb{P}\Big(\Lambda' \geq \frac{1}{\delta}\Big) \leq \delta,$$

which implies that $\mathbb{E}[\log \Lambda] \leq 1$. From (B.17),

$$\mathbb{E}[\zeta] \leq \frac{2(1 + 1/\log(n)) + cd\log(\log(d\log(n)))}{\varepsilon_n \beta_n^2}. \tag{B.18}$$

From (B.18), we have

$$\limsup_{n\to\infty} \frac{I_2}{\log(n)} \le \limsup_{n\to\infty} \frac{\Delta_{\max}\mathbb{E}[\zeta]}{\log(n)} = 0, \tag{B.19}$$

since β_n and ε_n are both sub-logarithmic. It remains to bound I_1. When $s(t) \ge \zeta$, from the definition of ζ in (B.15) we have

$$\langle x, \widehat{\theta}_t\rangle - \langle x, \theta\rangle \le \beta_n,$$

holds for any $x \in \mathcal{A}$. For each $m \in [M]$, we have

$$\begin{aligned}
\widehat{\Delta}_{x_m^*}(t) &= \langle \widehat{\theta}_t, \widehat{x}_m^*(t)\rangle - \langle \widehat{\theta}_t, x_m^*\rangle \\
&= \langle \widehat{\theta}_t, \widehat{x}_m^*(t)\rangle - \langle \theta, \widehat{x}_m^*(t)\rangle - \langle \widehat{\theta}_t, x_m^*\rangle + \langle \theta, x_m^*\rangle - \langle \theta, x_m^*\rangle + \langle \theta, \widehat{x}_m^*(t)\rangle \\
&\le 2\beta_t - \Delta_{\min}.
\end{aligned}$$

When n is sufficiently large, it holds that $\beta_n \le \Delta_{\min}/2$. This implies $\widehat{\Delta}_{x_m^*}(t) = 0$ such that $x_m^* = \widehat{x}_m^*(t)$ for all $t : s(t) > \zeta$. For notation simplicity, we denote $\mathcal{E}_t = \mathcal{B}_t^c \cap \mathcal{D}_{t,c_t}^c \cap \{s(t) \ge \zeta\} \cap \{c_t \in \mathcal{M}_t\}$. When $\mathcal{E}_t$ occurs, the algorithm is in the unwasted exploration stage and $x_m^* = \widehat{x}_m^*(n)$.

When $\mathcal{D}_{t,c_t}^c$ occurs and $c_t \in \mathcal{M}_t$, there exists $x' \in \mathcal{A}^{c_t}$ such that $N_{x'}(t) \le \min(f_n/\widehat{\Delta}^2_{\min}(t), T_{x'}(\widehat{\Delta}(t)))$. From the design of Algorithm 1, it holds that

- If $x = b_1$, then $N_x(t) \le \min(f_n/\widehat{\Delta}^2_{\min}(t), T_x(\widehat{\Delta}(t)))$.
- If $x = b_2$, then $N_x(t) = \min_{x\in\mathcal{A}^{c_t}} N_x(t) \le \min(f_n/\widehat{\Delta}^2_{\min}(t), T_{x'}(\widehat{\Delta}(t)))$.

Since the algorithm either pulls b_1 or b_2 in the unwasted exploration, it implies an upper bound for $s(t)$:

$$s(t) \le \sum_{x\in\mathcal{A}^{c_t}} N_x(t) \le |\mathcal{A}| \max_x \min(f_n/\widehat{\Delta}^2_{\min}(t), T_x(\widehat{\Delta}(t))). \tag{B.20}$$

Let Λ be the random variable given by

$$\Lambda = \min\left\{\lambda : \max_{x\in\mathcal{A}} |\langle x, \widehat{\theta}_t - \theta\rangle| \le \|x\|_{G_t^{-1}} f_{n,1/\lambda}^{1/2} \text{ for all } t \in [n]\right\},$$

where $f_{n,1/\lambda}$ is defined in Eq. (A.11). By the concentration inequality Lemma A.2, for any $\lambda \ge 1$,

$$\mathbb{P}\left(\Lambda \ge \lambda\right) \le 1/\lambda\,. \tag{B.21}$$

Hence the event $F = \{\Lambda \ge n\}$ satisfies $\mathbb{P}(F) \le 1/n$. Denote $\alpha_x^m(\Delta) = T_x^m(\Delta)/f_n$ where $T_x^m(\Delta)$ is the solution of optimisation problem in Definition 4.1 with true Δ. Given $\upsilon > 0$ let

$$\upsilon(\delta) = \sup\left\{\|\alpha(\Delta) - \alpha(\widetilde{\Delta})\|_\infty : \|\widetilde{\Delta} - \Delta\|_\infty \le \delta\right\},$$

where $\alpha(\Delta) = \{\alpha_x^m(\Delta)\}_{x\in\mathcal{A}^m, m\in[M]}$. By continuity assumption of α at Δ we have $\lim_{\delta\to 0}\upsilon(\delta) = 0$. Moreover, let's define

$$\tau_\delta = \min\left\{t : \max_{x\in\mathcal{A}} |\langle x, \widehat{\theta}_s - \theta\rangle| \le \delta/2 \text{ for all } x \in \mathcal{A} \text{ and } s \ge t\right\}.$$

Since $N_x(t) \ge \varepsilon_n s(t)/2$,

$$\max_{x\in\mathcal{A}} |\langle x, \widehat{\theta}_t - \theta\rangle| \le \sqrt{\frac{2f_{n,\Lambda}}{\varepsilon_n s(t)}}\,.$$

Therefore the number of exploration steps at time τ_δ is bounded by $s(\tau_\delta) \leq 8 f_{n,1/\Lambda} \varepsilon_n^{-1} \delta^{-2}$.

Let $(\delta_n)_{n=1}^{\infty}$ be a sequence with $\lim_{n\to\infty} \delta_n = 0$ and $\log(\log(n))/\delta_n^2 = o(\log(n))$. I_{11} decomposed as

$$\begin{aligned} I_{11} &= \mathbb{E}\Big[\sum_{t=1}^{n} \sum_{x \in \mathcal{A}_{-}^{c_t}} \Delta_x \mathbb{I}(X_t = x, \mathcal{E}_t)\Big] \\ &\leq \mathbb{E}\left[s(\tau_{\delta_n})\right] + \mathbb{E}\Big[\sum_{t=\tau_{\delta_n}}^{n} \sum_{x \in \mathcal{A}_{-}^{c_t}} \Delta_x \mathbb{I}(X_t = x, \mathcal{E}_t)\Big]. \end{aligned} \tag{B.22}$$

The first term in (B.22) is bounded by

$$\mathbb{E}\left[s(\tau_{\delta_n})\right] \leq \frac{8}{\varepsilon_n \delta_n^2} \mathbb{E}[f_{n,1/\Lambda}] = o(\log(n))\,,$$

where we used the assumption on (δ_n) and the fact that $\mathbb{E}[f_{n,1/\Lambda}] = O(\log\log(n))$. By the continuity assumption, the following statement holds

$$\begin{aligned} \sum_{t=\tau_{\delta_n}+1}^{n} \mathbb{I}(X_t = x, \mathcal{E}_t) &\leq \varepsilon_n s(n) + f_n \min\Big(1/\widehat{\Delta}_{\min}^2(n), \alpha_x^{c_t}(\widehat{\Delta}(n))/2\Big) \\ &\leq \varepsilon_n s(n) + f_n \min\Big(\frac{1}{\widehat{\Delta}_{\min}^2(n)}, (\alpha_x^{c_t}(\Delta) + \upsilon(\delta_n))/2\Big). \end{aligned} \tag{B.23}$$

The second term in (B.22) is bounded by

$$\begin{aligned} &\mathbb{E}\Big[\sum_{t=\tau_{\delta_n}}^{n} \sum_{x \in \mathcal{A}_{-}^{c_t}} \Delta_x \mathbb{I}(X_t = x, \mathcal{E}_t)\Big] \\ \leq\ &\mathbb{E}\Big[\sum_{m=1}^{M} \sum_{x \in \mathcal{A}_{-}^{m}} \Delta_x \sum_{t=1}^{n} \mathbb{I}(X_t = x, \mathcal{E}_t)\Big] \\ \leq\ &\mathbb{E}\Big[\sum_{m=1}^{M} \sum_{x \in \mathcal{A}_{-}^{m}} \Delta_x \varepsilon_n s(n) \mathbb{I}(\mathcal{E}_n)\Big] + \mathbb{E}\Big[\sum_{m=1}^{M} \sum_{x \in \mathcal{A}_{-}^{m}} \Delta_x f_n (\alpha_x^m(\Delta) + \upsilon(\delta_n))/2\mathbb{I}(\mathcal{E}_n)\Big]. \end{aligned}$$

To bound the second term, we take the limit as n tends to infinity and the fact that $\lim_{n\to\infty} \upsilon(\delta_n) = 0$ and $f_n \sim 2\log(n)$ shows that

$$\limsup_{n\to\infty} \frac{1}{\log(n)} \mathbb{E}\Big[\sum_{m=1}^{M} \sum_{x \in \mathcal{A}_{-}^{m}} \Delta_x f_n (\alpha_x^m(\Delta) + \upsilon(\delta_n))/2\mathbb{I}(\mathcal{E}_n)\Big] \leq \mathcal{C}(\theta, \mathcal{A}^1, \ldots, \mathcal{A}^M). \tag{B.24}$$

We bound the first term in the following lemma. The detailed proofs are deferred to Section B.4.

Lemma B.3. The regret contributed by the forced exploration satisfies

$$\limsup_{n\to\infty} \frac{\mathbb{E}\Big[\sum_{x \in \mathcal{A}_{-}^{c_t}} \Delta_x \varepsilon_n s(n) \mathbb{I}(\mathcal{E}_n)\Big]}{\log(n)} = 0.$$

This ends the proof. ■

B.4 Proof of Lemma B.3: Forced Exploration Regret

By the upper bound of unwasted exploration counter $s(n)$ in (B.20), it holds that

$$\begin{aligned}\sum_{m=1}^{M}\sum_{x\in\mathcal{A}_-^m}\Delta_x^m\varepsilon_n s(n)\mathbb{I}(\mathcal{E}_n) &\leq \sum_{m=1}^{M}\sum_{x\in\mathcal{A}_-^m}\Delta_x^m\varepsilon_n|\mathcal{A}|\max_x\min(f_n/\widehat{\Delta}_{\min}^2(n),T_x(\widehat{\Delta}(n)))\mathbb{I}(\mathcal{E}_n)\\ &\leq \varepsilon_n|\mathcal{A}|\sum_{m=1}^{M}\sum_{x\in\mathcal{A}_-^m}\Delta_x^m f_n/\widehat{\Delta}_{\min}(n)\mathbb{I}(\mathcal{E}_n).\end{aligned}$$

When event $\mathcal{E}_n$ occurs,

$$\begin{aligned}\max_{x\neq\widehat{x}_m^*(n)}\frac{(\Delta_x^m)^2}{(\widehat{\Delta}_x(n))^2} &\leq \max_{x\neq\widehat{x}_m^*(n)}\frac{(\Delta_x^m)^2}{(\Delta_x^m-2\beta_n)^2}\\ &= \max_{x\neq\widehat{x}_m^*(n)}\Big(1+\frac{4(\Delta_x^m-\beta_n)\beta_n}{(\Delta_x^m-2\beta_n)^2}\Big)\leq 1+\frac{16\beta_n}{\Delta_{\min}},\end{aligned}$$

For any $x\in\mathcal{A}^m$,

$$\widehat{\Delta}_{\min}(n)\geq\frac{1}{1+16\beta_n/\Delta_{\min}}\Delta_{\min}. \tag{B.25}$$

Since $\varepsilon_n=1/(\log\log(n))$, we have

$$\limsup_{n\to\infty}\frac{\sum_{x\in\mathcal{A}_-}\Delta_x\varepsilon_n s(n)\mathbb{I}(\mathcal{E})}{\log(n)}=0. \tag{B.26}$$

This ends the proof. ■

B.5 Proof of Lemma B.2: Wasted Exploration

First, we define

$$\mathcal{F}_s=\Big\{\exists t\geq d,\exists x:\langle x,\widehat{\theta}_t\rangle-\langle x,\theta\rangle\geq\|x\|_{G_t^{-1}}f_{n,1/s^2}^{1/2}\Big\}, \tag{B.27}$$

where $f_{n,1/s^2}$ is defined in Lemma A.2. From Lemma A.2, we also have $\mathbb{P}(\mathcal{F}_s)\leq 1/s^2$. Let $s'(t),s(t)$ be the number of rounds for wasted explorations, unwasted explorations until round t accordingly, and x_t^* is the optimal arm at round t. We decompose the regret as follows

$$R_{\text{we}}\leq\underbrace{\mathbb{E}\Big[\sum_{t\in\text{wasted}}\mathbb{I}(\mathcal{F}_{s'(t)})\Delta_{\max}\Big]}_{I_1}+\underbrace{\mathbb{E}\Big[\sum_{t\in\text{unwasted}}\mathbb{I}(\mathcal{F}_{s'(t)}^c)\langle x_t^*-X_t,\theta\rangle\Big]}_{I_2}. \tag{B.28}$$

To bound I_1, we have

$$I_1\leq\sum_{s=1}^{n}\mathbb{P}(\mathcal{F}_s)\Delta_{\max}\leq\sum_{s=1}^{n}\frac{1}{s^2}\Delta_{\max}=(2-\frac{1}{n})\Delta_{\max}. \tag{B.29}$$

To bound I_2, let's denote $\widetilde{\theta}_t$ as the optimistic estimator. Following the standard one step regret decomposition (See the proof of Theorem 19.2 in Lattimore and Szepesvári (2019) for details), it holds that

$$\begin{aligned}\langle x_t^*-X_t,\theta\rangle &= \langle x_t^*,\theta\rangle-\langle X_t,\theta\rangle\\ &\leq \langle X_t,\widetilde{\theta}_t\rangle-\langle X_t,\theta\rangle\\ &= \langle X_t,\widehat{\theta}_t-\theta\rangle+\langle X_t,\widetilde{\theta}_t-\widehat{\theta}_t\rangle.\end{aligned}$$

When $\mathcal{F}^c_{s'(t)}$ occurs, we have

$$\langle X_t, \widehat{\theta}_t - \theta\rangle \leq \|X_t\|_{G_t^{-1}} f_{n,1/(s'(t)^2)}, \langle X_t, \widetilde{\theta}_t - \widehat{\theta}_t\rangle \leq \|X_t\|_{G_t^{-1}} f_{n,1/(s'(t)^2)}.$$

Putting the above results together, we have

$$\langle x_t^* - X_t, \theta\rangle \leq 2\|X_t\|_{G_t^{-1}} f^{1/2}_{n,1/(s'(t)^2)}.$$

Applying Lemma C.3, we can bound I_2 as follows

$$\begin{aligned} I_2 &\leq \mathbb{E}\Big[2f^{1/2}_{n,1/(s'(t)^2)} \sum_{t\in\text{wasted}} \|X_t\|_{G_t^{-1}}\Big] \\ &\leq \mathbb{E}\Big[2f^{1/2}_{n,1/(s'(t)^2)} \sqrt{2s'(n)d\log\Big(\frac{s'(n)+d}{d}\Big)}\Big]. \end{aligned} \tag{B.30}$$

Recall that $p_{\min} = \min_m p_m$ be the minimum probability that each action set arrives. It is easy to see $\mathbb{P}(c_t \in \mathcal{M}_t | \mathcal{D}^c_{t,c_t}) = \mathbb{P}(c_t \in \mathcal{M}_t | \mathcal{M}_t \neq \emptyset) = \sum_{m\in\mathcal{M}_t} p_m \geq p_{\min}$. We bound $s'(n)$ by $s(n)$ as follows

$$\begin{aligned} \mathbb{E}[s'(n)] = \mathbb{E}\Big[\sum_{t=1}^n \mathbb{I}\Big(\mathcal{D}^c_{t,c_t}, c_t \notin \mathcal{M}_t\Big)\Big] = \sum_{t=1}^n \mathbb{P}(\mathcal{D}^c_{t,c_t})\mathbb{P}(c_t \notin \mathcal{M}_t|\mathcal{D}^c_{t,c_t}) \\ \leq \frac{1}{p_{\min}} \sum_{t=1}^n \mathbb{P}(\mathcal{D}^c_{t,c_t})\mathbb{P}(c_t \in \mathcal{M}_t|\mathcal{D}^c_{t,c_t}) \\ = \frac{1}{p_{\min}}\mathbb{E}\Big[\sum_{t=1}^n \mathbb{I}\Big(\mathcal{D}^c_{t,c_t}, c_t \in \mathcal{M}_t\Big)\Big] = \frac{1}{p_{\min}}\mathbb{E}[s(n)]. \end{aligned} \tag{B.31}$$

Putting (B.29)-(B.31) together, The regret in the wasted exploration can be upper bounded by

$$R_{\text{we}} \leq (2 - \frac{1}{n})\Delta_{\max} + \frac{2}{p_{\min}}\sqrt{2d\log\Big(\frac{s(n)/p_{\min}+d}{d}\Big) f_{n,(p_{\min}/s(n))^2} s(n)/p_{\min}}, \tag{B.32}$$

where $f_{n,(p_{\min}/s(n))^2}$ is defined in (A.2).

Next, we recall the upper bound (B.20) for the number of pulls in unwasted exploration,

$$\begin{aligned} s(n) &\leq |\mathcal{A}| \max_x \min\Big\{f_n/\widehat{\Delta}_{\min}(n), T_x(\widehat{\Delta}(n))\Big\} \\ &\leq |\mathcal{A}| f_n/\widehat{\Delta}_{\min}(n). \end{aligned}$$

From (B.25), we have

$$\widehat{\Delta}_{\min}(n) \geq \frac{1}{1+\delta_n}\Delta_{\min} \geq \frac{\Delta^2_{\min}}{\Delta_{\min} + 16\beta_n},$$

where $\beta_n = 1/\log(\log(n))$. Overall, we see $s(n) \leq \mathcal{O}(\log(n))$. Plugging this into (B.32), we reach

$$\limsup_{n\to\infty} \frac{R_{\text{we}}}{\log(n)} = 0.$$

This ends the proof. ■

B.6 Proof of Lemma A.1

First, we start by the following claim:

Claim B.4. Assume H_n is a sequence of $d \times d$ positive definite matrices such that $H_n \to H$ and H is positive semidefinite. Then, $HH_n^{-1}H \to H$ as $n \to \infty$.

Proof. Without loss of generality, we can assume that H is given in the block matrix form

$$H = \begin{pmatrix} A & 0 \\ 0 & 0 \end{pmatrix}$$

where A is a nonsingular $m \times m$ matrix with $m > 0$. (If $m = 0$, H is the all zero matrix and the claim trivially holds.) Consider the same block partitioning of H_n:

$$H_n = \begin{pmatrix} A_n & B_n \\ B_n^\top & D_n \end{pmatrix},$$

where A_n is thus also an $m \times m$ matrix. Clearly, $A = \lim_{n\to\infty} A_n$ and A_n is nonsingular (or H_n would be singular), while $B_n \to B$ and $D_n \to D$ where all entries in B and D are zero. Then, as is well known,

$$H_n^{-1} = \begin{pmatrix} A_n^{-1} + A_n^{-1}B_nS_n^{-1}B_n^\top A_n^{-1} & -A_n^{-1}B_nS_n^{-1} \\ -S_n^{-1}B_n^\top A_n^{-1} & S_n^{-1} \end{pmatrix}.$$

where $S_n = D_n - B_n^\top A_n^{-1}B_n$ is the Schur-complement of block D_n of matrix H_n. Note that

$$HH_n^{-1}H = \begin{pmatrix} A(A_n^{-1} + A_n^{-1}B_nS_n^{-1}B_n^\top A_n^{-1})A & 0 \\ 0 & 0 \end{pmatrix}.$$

Since the matrix inverse is continuous if the limit is nonsingular, $A_n^{-1} \to A^{-1}$. Clearly, it suffices to show that $A_n^{-1} + A_n^{-1}B_nS_n^{-1}B_n^\top A_n^{-1} \to A^{-1}$. Hence, it remains to check that $A_n^{-1}B_nS_n^{-1}B_n^\top A_n^{-1} \to 0$. This follows because $B_n \to B$ and $D_n \to D$ and $S_n \to D - B^\top A^{-1}B = 0$ where $D = 0$ and $B = 0$. □

Proof of Lemma A.1. Let $L = \limsup_{n\to\infty} \log(n)\|s\|^2_{G_n^{-1}}$. We need to prove that $L \leq 1/c$. Without loss of generality, assume that $L > 0$ (otherwise there is nothing to be proven) and that for some H positive semidefinite matrix, $\zeta \in \mathbb{R}$ and $\kappa \in \mathbb{R} \cup \{\infty\}$, *(i)* $\log(n)\|s\|^2_{G_n^{-1}} \to L$; *(ii)* $H_n = G_n^{-1}/||G_n^{-1}|| \to H$; *(iii)* $\lambda_{\min}(G_n)/\log(n) \to \zeta > 0$ and *(iv)* $\frac{\rho_n(H)}{\log(n)\|s\|^2_{G_n^{-1}}} \to \kappa \geq c$. We claim that $\|s\|_H > 0$, hence $\rho_n(H)$ is well-defined and in particular $\rho_n(H) \to 1$ as $n \to \infty$. If this was true, then the proof was ready since

$$L = \lim_{n\to\infty} \frac{\log(n)\|s\|^2_{G_n^{-1}}}{\rho_n(H)} = \frac{1}{\lim_{n\to\infty} \frac{\rho_n(H)}{\log(n)\|s\|^2_{G_n^{-1}}}} = 1/\kappa \leq 1/c\,.$$

Hence, it remains to show the said claim. We start by showing that $\|s\|_H > 0$. For this note that $||G_n^{-1}|| = 1/\lambda_{\min}(G_n)$ and hence

$$\|s\|^2_{\frac{G_n^{-1}}{||G_n^{-1}||}} = \frac{\lambda_{\min}(G_n)}{\log(n)} \|s\|^2_{G_n^{-1}} \log(n)\,.$$

Taking the limit of both sides, we get $\|s\|^2_H \to \zeta L > 0$. Now,

$$\rho_n(H) = \frac{\|s\|^2_{G_n^{-1}}\|s\|^2_{HG_nH}}{\|s\|^4_H} = \frac{\|s\|^2_{H_n}\|s\|^2_{HH_n^{-1}H}}{\|s\|^4_H} \overset{n\to\infty}{\to} \frac{\|s\|^2_H\|s\|^2_H}{\|s\|^4_H} = 1\,,$$

where we used Claim B.4. □

B.7 Proof of Theorem 3.9

Suppose that $\{x_m^* : m \in [M]\}$ spans $\mathbb{R}^d$. Recall that LinUCB chooses

$$X_t = \operatorname*{argmax}_{x \in \mathcal{A}^{c_t}} \langle x, \widehat{\theta}_{t-1} \rangle + ||x||_{G_{t-1}^{-1}} \beta_t^{1/2} \,,$$

where $\beta_t = O(d \log(t))$ is chosen so that

$$\mathbb{P}\left(||\widehat{\theta}_t - \theta||_{G_t} \geq \beta_t\right) \leq 1/t^3 \,,$$

which is known to be possible (Lattimore and Szepesvári, 2019, §20). Define F_t to be the event that $||\widehat{\theta}_t - \theta||_{G_t} \geq \beta_t$. Then the instantaneous pseudo-regret of LinUCB is bounded by

$$\Delta_t \leq \mathbf{1}_{F_t} + \langle x_t^* - X_t, \theta \rangle \leq \mathbf{1}_{F_t} + 2\beta_t^{1/2} ||X_t||_{G_t^{-1}} \leq \mathbf{1}_{F_t} + 2\sqrt{\beta_t ||G_t^{-1}||} \,,$$

where the matrix norm is the operator name (in this case, maximum eigenvalue). Let $\tau = 1 + \max\{t : F_t \text{ holds}\}$, which satisfies $\mathbb{E}[\tau] = O(1)$. The cumulative regret after τ is bounded almost surely by

$$\sum_{t=\tau}^{n} \langle x_t^* - X_t, \theta \rangle = O\left(\sqrt{n} \log(n)\right) \,,$$

where the Big-Oh hides constants that only depend on the dimension. Hence all optimal arms are played linearly often after τ, which by the assumption that $\{x_m^* : m \in [M]\}$ spans $\mathbb{R}^d$ implies that $||G_t^{-1}|| = O(1/t)$. Hence the instantaneous regret for times $t \geq \tau$ satisfies

$$\Delta_t = O\left(\sqrt{\frac{\beta_t}{t}}\right) \,.$$

Since $\Delta_t \in \{0\} \cup [\Delta_{\min}, 1]$, it follows that the regret vanishes once $\Delta_t < \Delta_{\min}$. But by the previous argument and the assumption on β_t we have for $t \geq \tau$ that

$$\Delta_t \leq 2\sqrt{\beta_t ||G_t^{-1}||} = O\left(\sqrt{\frac{\log(t)}{t}}\right) \,.$$

Hence for sufficiently large t (independent of n) the regret vanishes, which completes the proof.

C Supporting Lemmas

Lemma C.1 (Bretagnolle-Huber Inequality)**.** Let $\mathbb{P}$ and $\widetilde{\mathbb{P}}$ be two probability measures on the same measurable space $(\Omega, \mathcal{F})$. Then for any event $\mathcal{D} \in \mathcal{F}$,

$$\mathbb{P}(\mathcal{D}) + \widetilde{\mathbb{P}}(\mathcal{D}^c) \geq \frac{1}{2} \exp\left(-\mathrm{KL}(\mathbb{P}, \widetilde{\mathbb{P}})\right) \,, \tag{C.1}$$

where $\mathcal{D}^c$ is the complement event of $\mathcal{D}$ ($\mathcal{D}^c = \Omega \setminus \mathcal{D}$) and $\mathrm{KL}(\mathbb{P}, \widetilde{\mathbb{P}})$ is the KL-divergence between $\mathbb{P}$ and $\widetilde{\mathbb{P}}$, which is defined as $+\infty$, if $\mathbb{P}$ is not absolutely continuous with respect to $\widetilde{\mathbb{P}}$, and is $\int_\Omega d\mathbb{P}(\omega) \log \frac{d\mathbb{P}}{d\widetilde{\mathbb{P}}}(\omega)$ otherwise.

The proof can be found in the book of Tsybakov (2008). When $\mathrm{KL}(\mathbb{P}, \widetilde{\mathbb{P}})$ is small, we may expect the probability measure $\mathbb{P}$ is close to the probability measure $\widetilde{\mathbb{P}}$. Note that $\mathbb{P}(\mathcal{D}) + \mathbb{P}(\mathcal{D}^c) = 1$. If $\widetilde{\mathbb{P}}$ is close to $\mathbb{P}$, we may expect $\mathbb{P}(\mathcal{D}) + \widetilde{\mathbb{P}}(\mathcal{D}^c)$ to be large.

Lemma C.2 (Divergence Decomposition)**.** Let $\mathbb{P}$ and $\widetilde{\mathbb{P}}$ be two probability measures on the sequence $(A_1, Y_1, \ldots, A_n, Y_n)$ for a fixed bandit policy π interacting with a linear contextual bandit with standard Gaussian noise and parameters θ and $\widetilde{\theta}$ respectively. Then the KL divergence of $\mathbb{P}$ and $\widetilde{\mathbb{P}}$ can be computed exactly and is given by

$$\mathrm{KL}(\mathbb{P}, \widetilde{\mathbb{P}}) = \frac{1}{2} \sum_{x \in \mathcal{A}} \mathbb{E}[T_x(n)] \, \langle x, \theta - \widetilde{\theta} \rangle^2 \,, \tag{C.2}$$

where $\mathbb{E}$ is the expectation operator induced by $\mathbb{P}$.

This lemma appeared as Lemma 15.1 in the book of Lattimore and Szepesvári (2019), where the reader can also find the proof.

Lemma C.3. Let $\{X_t\}_{t=1}^{\infty}$ be a sequence in $\mathbb{R}^d$ satisfying $\|X_t\|_2 \leq 1$ and $G_t = \sum_{s=1}^{t} X_t X_t^\top$. Suppose that $\lambda_{\min}(G_d) \geq c$ for some strictly positive c. For all $n > 0$, it holds that

$$\sum_{t=d+1}^{n} \|X_t\|_{G_t^{-1}} \leq \sqrt{2nd \log(\frac{d+n}{d})}.$$

Lemma C.4. Let $\varepsilon > 0$ and denote $T(\widehat{\Delta}(n)) \in \mathbb{R}^{|\mathcal{A}|}$ as the solution of the optimisation problem defined in Definition 4.1. Then we define

$$S_\varepsilon(\widehat{\Delta}(n)) = \min\Big\{\varepsilon f_n, T(\widehat{\Delta}(n))\Big\}.$$

Then for all $x \in \mathcal{A}$,

$$\|x\|^2_{H^{-1}_{S_\varepsilon(\widehat{\Delta}(n))}} \leq \max\Big\{\frac{\varepsilon^2}{f_n}, \frac{\widehat{\Delta}_x^2(n)}{f_n}\Big\}.$$

This is Lemma 17 in the book of Lattimore and Szepesvári (2017), where the reader can also find the proof.

Lemma C.5. Suppose that $T_x^m(\cdot)$ is uniquely defined at Δ. Then it is continuous at Δ.

Proof. Suppose it is not continuous. Then there exists a sequence $(\Delta_n)_{n=1}^{\infty}$ with $\lim_{n\to\infty} ||\Delta_n - \Delta|| = 0$ and for which $\lim_{n\to\infty} T_x^m(\Delta_n) \neq T_x^m(\Delta)$ for some m and $x \in \mathcal{A}^m$. Since $\Delta_n \to \Delta$ it follows that for sufficiently large n the optimal actions with respect to Δ_n are the same as Δ. Hence, for sufficiently large n, by the definition of the optimisation problem,

$$T^m_{x^*_m}(\Delta_n) = \infty = T^m_{x^*_m}(\Delta)\,.$$

Therefore there exists a context m and suboptimal action $x \neq x^*_m$ such that $\lim_{n\to\infty} T_x^m(\Delta_n) \neq T_x^m(\Delta)$. It is easy to check that the value of the optimisation problem is continuous. Specifically, that

$$\lim_{n\to\infty} \sum_{m=1}^{M} \sum_{x \in \mathcal{A}^m} T_x^m(\Delta_n) = \sum_{m=1}^{M} \sum_{x \in \mathcal{A}^m} T_x^m(\Delta)\,.$$

Hence $\limsup_{n\to\infty} T_x^m(\Delta_n) < \infty$ for $x \neq x^*_m$. Therefore a compactness argument shows there exists a cluster point S of the allocation $(T(\Delta_n))_{n=1}^{\infty}$ with $S_m^x \neq T_m^x(\Delta)$ for some m and $x \neq x^*_m$. And yet by the previous display

$$\sum_{m=1}^{M} \sum_{x \in \mathcal{A}^m} S_x^m = \sum_{m=1}^{M} \sum_{x \in \mathcal{A}^m} T_x^m(\Delta)\,.$$

Since the constraints of the optimisation problem are continuous it follows that S also satisfies the constraints in the optimisation problem and so $S \neq T(\Delta)$ is another optimal allocation, contradicting uniqueness. Therefore $T_x^m(\cdot)$ is continuous at Δ. □

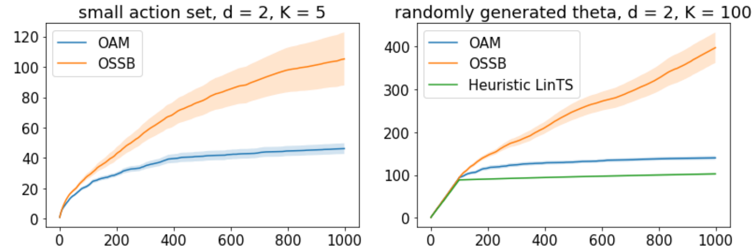

Figure 4: The left panel is for small size action set and the right panel is for randomly generated θ. The results are averaged over 100 realisations.

D Additional Experiments

In this section, we consider two more experiment settings in Figure 4.

1. Small size action set. We conduct the experiments with the number of action set equal to 5. Comparing with large size action set (Section 5.4), we found that OAM still outperforms OSSB but the improvement is smaller, as one might expect.

2. Randomly generated θ. For each replication, θ is randomly generated from multivariate normal with variance 10 and we normalise θ such that its ℓ_2 norm is 1. OAM still outperforms OSSB for randomly generated θ. In addition, we compare with the heuristic LinTS (remove all the variance blowup factors and use a Gaussian prior). We find that the heuristic LinTS enjoys the best performance by a modest margin. Analysing heuristic LinTS, however, remains a fascinating open problem. As far as we are aware, it is not known whether or not it even achieves sublinear regret in the worst case.