A PROOF OF THEOREM 3.1

Theorem 3.1. *Consider a K-armed utility-based two-dueling bandits game. Assume that the BAIM S in DoublerBAI has a sample complexity of $O(H\ln(\frac{H}{\delta}))$, where S outputs the best arm with probability at least $1-\delta$. Given an exponentially growing sequence $\{T_i\}_{i\in\mathbb{N}}$ with parameters $a, b > 1$, i.e., $T_i = \lfloor a^{b^i} \rfloor$, the expected regret of DoublerBAI is bounded by*

$$\begin{aligned}\mathbb{E}[R_T] =& O((H\ln H)^b) + O(H\ln T)\\ &+ O(H\ln H\ln\ln T) + O(\ln\ln T),\end{aligned}$$

where $H := \sum_{i=2}^{K} \frac{1}{\Delta_i^2}$ is the problem complexity for a bandit instance.

Proof. For convenience, we recall the notation from the proof sketch of Theorem 3.1. $B(\delta)$ denotes the supremum of the expected regret of the BAIM S to identify the best arm with probability at least $1-\delta$. We also denote the sample complexity of the BAIM S to identify the best arm with probability at least $1-\delta$ by $\widetilde{B}(\delta)$. Because each sample on arm $x_i \in \mathcal{X}$ incurs regret $\Delta_i \le 1$, $B(\delta)$ and $\widetilde{B}(\delta)$ have the same order of $O(H\ln(\frac{H}{\delta}))$ [13]. Thus, we can assume $B(\delta) = c_1 H\ln(\frac{H}{\delta})$ and $\widetilde{B}(\delta) = c_2 H\ln(\frac{H}{\delta})$, where $c_2 \ge c_1 \ge 1$. Let R_t^{left} and R_t^{right} denote the regret incurred by the left and right arm of the played two arms (x_t, y_t) at time-step t, respectively .

We firstly consider R_t^{right} of the exploration stage. From the definition of regret in the dueling bandits problem, we have that

$$\begin{aligned}&\mathbb{E}[\sum_{t=1}^{\tau_i^{\text{explore}}} R_t^{\text{right}}]\\ =&\mathbb{E}[\sum_{t=1}^{\tau_i^{\text{explore}}} \frac{\mu(x_1)-\mu(y_t)}{2}]\\ =&\mathbb{E}[\sum_{t=1}^{\tau_i^{\text{explore}}} (\frac{\mu(x_1)-\mu(\bar{x}_i)+1}{2} - \frac{\mu(y_t)-\mu(\bar{x}_i)+1}{2})] \qquad (1)\\ \le& B(\frac{1}{\tau_{i+1}}). \qquad (2)\end{aligned}$$

The last inequality holds because Eq. (1) is the regret of the BAI game in epoch i, and Eq. (2) is a upper bound of that regret.

In the stage of exploitation, the right arm is simply chosen as the identified best arm of the exploration stage, with error probability at most $\frac{1}{\tau_{i+1}}$.

Thus, we can bound $\mathbb{E}[\sum_{t=1}^{\tau_i} R_t^{\text{right}}]$ by

$$\begin{aligned}\mathbb{E}[\sum_{t=1}^{\tau_i} R_t^{\text{right}}] &\le (1-\frac{1}{\tau_{i+1}})B(\frac{1}{\tau_{i+1}}) + \frac{1}{\tau_{i+1}}O(\tau_i)\\ &\le B(\frac{1}{\tau_{i+1}}) + O(1).\end{aligned}$$

Now we consider R_t^{left}. In epoch i, the left arm $\bar{x}_i$ is chosen in two cases: (i) $\bar{x}_i$ is chosen as the identified best arm $\hat{x}_{i-1}$ of epoch $i-1$ when S terminates and returns $\hat{x}_{i-1}$ in epoch $i-1$, and (ii) $\bar{x}_i$ is randomly chosen from $\mathcal{X}$ otherwise.

In case (i), the probability of $\hat{x}_{i-1} \ne x_1$ is at most $\frac{1}{\tau_i}$, and thus we have

$$\mathbb{E}[\sum_{t=1}^{\tau_i} R_t^{\text{left}}] \le \frac{1}{\tau_i}\cdot O(\tau_i) + (1-\frac{1}{\tau_i})\cdot 0 = O(1).$$

In case (ii), we simply have

$$\mathbb{E}[\sum_{t=1}^{\tau_i} R_t^{\text{left}}] = O(\tau_i).$$

Thus, in case (i), the regret is upper bounded by $B(\frac{1}{\tau_{i+1}}) + O(1)$. In case (ii), the regret is upper bounded by $O(\tau_i)$.

Below we bound the regret in case (i) and (ii), respectively.

First consider case (i), where we bound the regret using similar techniques of doubling trick [5, 6]. Let L_T denote the number of epochs up to time T. According to the definition of $\{T_i\}_{i\in\mathbb{N}}$ (*i.e.*, $T_i = \lfloor a^{b^i} \rfloor,\ a, b > 1$), we have that $\forall L_T > 1$,

$$L_T - 1 = \lceil \ln_b \ln_a T \rceil \le \ln_b \ln_a T + 1. \qquad (3)$$

Asymptotically for i and $T \to \infty$, $T_i = O(a^{b^i})$ and $L_T = O(\ln\ln T)$.

The expected regret up to time T in case (a) can be bounded by (the index of epoch i starts from 0)

$$\begin{aligned}&\sum_{i=0}^{L_T-1} \big(B(\frac{1}{\tau_{i+1}}) + O(1)\big)\\ \le& \sum_{i=0}^{L_T} (c_1 H\ln(H\tau_i) + O(1))\\ \le& \sum_{i=0}^{L_T} (c_1 H\ln(Ha^{b^i}) + O(1))\\ =& \sum_{i=0}^{L_T} (c_1 H\ln H + c_1 H\ln(a) b^i + O(1))\\ =& O(H\ln H\ln\ln T) + O(\ln\ln T)\\ &+ c_1 H\ln a \sum_{i=0}^{L_T} b^i, \qquad (4)\end{aligned}$$

where $\sum_{i=0}^{L_T} b^i$ can be bounded by $\sum_{i=0}^{L_T} b^i \le \frac{b}{b-1} b^{L_T} \le \frac{b^3}{b-1}\ln_a T = \frac{b^3}{b-1}\frac{\ln T}{\ln a}$ (using Eq. (3)).

Thus, we can bound Eq. (4) by

$$\begin{aligned}Eq.\ (4) \le& O(H\ln H\ln\ln T) + O(\ln\ln T)\\ &+ c_1 H\frac{b^3}{b-1}\ln T\\ =& O(H\ln H\ln\ln T) + O(\ln\ln T)\\ &+ O(H\ln T). \qquad (5)\end{aligned}$$

Then, we consider case (ii), where we will prove the regret is independent of T. In epoch i, with probability at least $1-\frac{1}{\tau_{i+1}}$, S will terminate and return the identified best arm $\hat{x}_i$ after $\widetilde{B}(\frac{1}{\tau_{i+1}}) = c_2 H\ln(H\tau_{i+1})$ time-steps. For ease of analysis, we regard the scenario where S does not terminate after $c_2 H\ln(H\tau_{i+1})$ time-steps as the scenario where S terminates but returns a wrong arm and our analysis still holds.

Note that if

$$c_2 H \ln(H\tau_{i+1}) \le \tau_i \tag{6}$$

holds, S will terminate and return $\hat{x}_i$, and thus case (ii) in epoch $i+1$ will not occur. For large enough τ_i, $c_2 H \ln(H\tau_{i+1}) \le \tau_i$ must holds. This is because that the left side grows logarithmically with τ_i and the right side grows linearly (τ_i is defined to grow exponentially). Thus, case (ii) only occurs in early epoch, for which the previous epoch does not satisfy this condition.

For $i > 0$, Eq. (6) can be written as

$$c_2 H \ln H + c_2 H \ln(T_{i+1} - T_i) \le T_i - T_{i-1}.$$

Notice that $T_{i+1} - T_i \le T_{i+1} \le (T_i+1)^b$, and similarly, $T_{i-1} \le (T_i+1)^{\frac{1}{b}}$. Thus, $\forall T_i \ge T_0 := \left\lceil (\frac{1}{2})^{\frac{b}{1-b}} - 1 \right\rceil$, $T_i - T_{i-1} \ge T_i - (T_i+1)^{\frac{1}{b}} \ge \frac{1}{2}T_i - \frac{1}{2}$.

Then we know that for any epoch i such that $T_i \ge \widetilde{T} = \max\{T_0, \lceil 20 c_2 b H \ln(c_2 b H) - 1\rceil\}$, S will terminate and return the identified best arm $\hat{x}_i$.

This is because that

$$c_2 H \ln H + c_2 H b \ln(T_i + 1) \overset{a}{\le} \frac{1}{2}T_i - \frac{1}{2} \overset{a}{\le} T_i - T_{i-1}$$

where (a) uses $T_i \ge \lceil 20 c_2 b H \ln(c_2 b H) - 1\rceil$ and (b) uses $T_i \ge T_0$.

Now we can obtain $L - 1 \le \left\lfloor \ln_b \ln_a \widetilde{T} \right\rfloor \le \ln_b \ln_a \widetilde{T} + 1$, and thus $T_{L-1} \le a^{b^{\ln_b \ln_a \widetilde{T}+1}} = \widetilde{T}^b$, where L is the number of the epochs in which case (ii) occurs, *i.e.*, case (ii) occurs in epoch $0, 1, \ldots, L-1$ and Eq. (6) firstly holds in epoch $L-1$.

Therefore, we can bound the expected regret in case (ii) by

$$\begin{aligned}
\mathbb{E}[\sum_{i=0}^{L-1} R^i] &= \sum_{i=0}^{L-1} O(\tau_i) \\
&= O(T_{L-1}) \\
&\le O(\widetilde{T}^b) \\
&= O\big((H \ln H)^b\big).
\end{aligned} \tag{7}$$

Summing up the expected regret in case (i) (Eq. (5)) and (ii) (Eq. (7)), we obtain the result of Theorem 3.1.

□

B PROOF OF THEOREM 3.3

In order to prove Theorem 3.3, we firstly introduce the following lemma.

Lemma 7.1. *When running the implement of the SBM (Algorithm 3) with $\alpha > 0$, the number of times a suboptimal arm x_i has been pulled up to time T, which is denoted by $\rho_i(T)$, satisfies*

$$\forall s \ge \frac{4(\alpha+4)}{\Delta_i^2} \ln T,\ \Pr[\rho_i(T) \ge s] < \frac{4}{\alpha}\Big(\frac{s}{2}\Big)^{-\alpha}.$$

Proof. Our analysis follows similar techniques as that in [3]. For ease of notation, we define $\beta := \alpha + 2$ and $u_i(t) := \frac{2\beta \ln t}{\Delta_i^2}$. Recall the notation defined in Algorithm 3. $\rho_i(t)$ denotes the number of times arm $x_i \in \mathcal{X}$ has been pulled up to time t and $s_i(t)$ denotes the number of times arm x_i has received the additional feedback up to time t.

At time-step t, if a suboptimal arm x_i was chosen, one of the following three events must be true.

$$\mathcal{E}_t := \{\rho_i(t) + s_i(t) < u_i(t)\}$$

$$\mathcal{F}_t := \{\hat{\mu}_i \ge \mu_i + \sqrt{\frac{\beta \ln t}{2(\rho_i(t) + s_i(t))}}\}$$

$$\mathcal{G}_t := \{\hat{\mu}_1 + \sqrt{\frac{\beta \ln t}{2(\rho_1(t) + s_1(t))}} \le \mu_1\}$$

If all three are false, we have

$$\begin{aligned}
&\hat{\mu}_1 + \sqrt{\frac{\beta \ln t}{2(\rho_i(t) + s_i(t))}} \\
>&\mu_1 \\
=&\mu_i + \Delta_i \\
\ge&\mu_i + 2\sqrt{\frac{\beta \ln t}{2(\rho_i(t) + s_i(t))}} \\
>&\hat{\mu}_i + \sqrt{\frac{\beta \ln t}{2(\rho_i(t) + s_i(t))}},
\end{aligned}$$

and then arm i cannot be chosen.

When $\rho_i(t) \ge u_i(T)$, event $\mathcal{E}_t$ is false. Thus, we have

$$\mathbb{E}[\rho_i(T) - u_i(T)] \le \sum_{t=u_i(T)+1}^{T} \Pr[\mathcal{F}_t \vee \mathcal{G}_t].$$

Using the Chernoff-Hoeffding bound, we can bound the probability of event $\mathcal{F}_t$ occurring by

$$\Pr[\mathcal{F}_t] \le \Pr\left[\exists(\rho_i(t) + s_i(t)) \in [2t] : \hat{\mu}_i \ge \mu_i + \sqrt{\frac{\beta \ln t}{2(\rho_i(t) + s_i(t))}}\right] \le 2t \cdot t^{-\beta} = 2t^{1-\beta}.$$

Analogously, $\Pr[\mathcal{G}_t] \le 2t^{1-\beta}$. Thus, we have

$$\begin{aligned}
\mathbb{E}[\rho_i(T) - u_i(T)] &\le \sum_{t=u_i(T)+1}^{T} 2 \cdot 2t^{1-\beta} \\
&\le \frac{4}{\beta - 2}\Big(\frac{2\beta \ln T}{\Delta_i^2}\Big)^{2-\beta}.
\end{aligned}$$

Let $\rho_i^s(T)$ denote the number of times arm x_i has been pulled between time s and T. For $s \ge \frac{2\beta \ln T}{\Delta_i^2}$, we have

$$\mathbb{E}[\rho_i^s(T) - u_i(T)] \le \sum_{t=s}^{T} 4t^{1-\beta} \le \frac{4}{\beta - 2} s^{2-\beta}.$$

Assuming that arm x_i has been pulled at least $s \ge \frac{4(\beta+2)\ln T}{\Delta_i^2}$ times up to time T, we have $\rho_i^{s-u_i(T)-1}(T) \ge u_i(T) + 1$. Thus, we can bound $\Pr[\rho_i(T) > s]$ by

$$\begin{aligned}
\Pr[\rho_i(T) \ge s] &\le \Pr[\rho_i^{s-u_i(T)-1}(T) - u_i(T) \ge 1] \\
&\overset{(a)}{\le} \mathbb{E}[\rho_i^{s-u_i(T)-1}(T) - u_i(T)]
\end{aligned}$$

$$\leq \frac{4}{\beta-2}(s-u_i(T)-1)^{2-\beta}$$
$$\overset{(b)}{\leq} \frac{4}{\beta-2}\Big(\frac{s}{2}\Big)^{2-\beta},$$

where (a) uses Markov's inequality and (b) uses $s \geq \frac{4(\beta+2)\ln T}{\Delta_i^2} \geq 2u_i(T)+2$.

Since $\beta := \alpha + 2$, we obtain the result of Lemma 7.1.

□

From Lemma 7.1, we know that our SBM's implement (Algorithm 3) also satisfies the α-robustness defined in [3], with the constant factor of the probability slightly enlarged.

Therefore, the analysis in (Theorem 4.2 in [3]) still holds. Let $\tau_{xy}(\tau_x(T))$ denote the number of times SBM S_x ($x \in \mathcal{X} \setminus \{x_1\}$) has advanced suboptimal arm y up to time T. We have that with parameter $\alpha = \max\{3, \frac{\ln K}{\ln\ln T}\}$, $\mathbb{E}[\tau_{xy}(\tau_x(T))]$ is bounded by

$$\mathbb{E}[\tau_{xy}(\tau_x(T))] = O\Big(\frac{\alpha}{\Delta_y^2}\big(\ln\ln T + \ln K + \ln(\frac{1}{\Delta_x})\big)\Big). \tag{8}$$

This conclusion will be used in the following proof of Theorem 3.3.

Theorem 3.3. *Consider a K-armed utility-based two-dueling bandits game. The expected regret of MultiSBM-Feedback, which implements an SBM defined in Algorithm 3, is bounded by*

$$\mathbb{E}[R_T] \leq \min\left\{\sum_{i>1}\frac{(\alpha+2)\Delta_{max}}{\Delta_i^2}\ln T,\ \sum_{i>1}\frac{2(\alpha+2)}{\Delta_i}\ln T\right\} + \frac{(\alpha+8)\Delta_{max}}{2\alpha}K + \sum_{j>1}\sum_{i>1}O\Big(\frac{\alpha\Delta_{max}}{\Delta_j^2}\big(\ln\ln T + \ln K + \ln(\frac{1}{\Delta_i})\big)\Big),$$

where $\Delta_{max} := \max_{i>1}\Delta_i$ *and the confidence interval parameter* $\alpha = \max\{3, \frac{\ln K}{\ln\ln T}\}$.

Proof. Let $\tau_x(t)$ denote the number of times SBM S_x ($x \in \mathcal{X}$) has been queried up to time t. Let $R_x(t')$ denote the regret seen by S_x up to its internal time t'. Let $R_{xy}(\tau_x(t))$ denote the regret due to S_x advancing suboptimal arm y up to time t.

In MultiSBM-Feedback, the right arm in each time-step equals to the left arm in the next time-step. Thus, we have that the expected regret $\mathbb{E}[R_T]$ up to time T can be bounded by

$$\mathbb{E}[R_T] \leq 0.5 + \mathbb{E}[R_{x_1}(T)] + \mathbb{E}[\sum_{y\neq x_1}\sum_{x\neq x_1}R_{xy}(\tau_x(T))].$$

The analysis of $\mathbb{E}[\sum_{y\neq x_1}\sum_{x\neq x_1}R_{xy}(\tau_x(T))]$ follows the similar line of [3]. In the following we focus on $\mathbb{E}[R_{x_1}(T)]$.

We inherit the notation and reasoning in the proof of Lemma 7.1. In MultiSBM-Feedback, up to any time t, the number of times arm x_1 being the left arm equals to the number of times arm x_1 being the right arm. Subtracting the number of times (x_1, x_1) being played for both side, we have that the number of times S_{x_1} advancing suboptimal arms equals to the number of times S_{x_1} receives additional feedback from suboptimal arms. Thus, in SBM S_{x_1},

$$\forall t,\ \sum_{i>1}\rho_i(t) = \sum_{i>1}s_i(t). \tag{9}$$

At time-step t, we denote the arm being pulled by I_t and the arm being observed through GetAdditionalFeedback by A_t. We also define $\{\Pi(x)\}$ to be the indicator function of the event $\Pi(x)$ for any predicate $\Pi(x)$. In SBM S_{x_1}, up to its internal time T, for any suboptimal arm x_i, we have

$$\begin{aligned}
&\rho_i(T) + s_i(T)\\
=&\sum_{t=1}^{T}\{I_t = x_i\} + \sum_{t=1}^{T}\{A_t = x_i\}\\
\leq& u_i(T) + \sum_{t=t_0+1}^{T}\{I_t = x_i, \rho_i(t)+s_i(t) \geq u_i(T)\}\\
&+ \sum_{t=t_0+1}^{T}\{A_t = x_i, \rho_i(t)+s_i(t) \geq u_i(T)\}\\
\leq& u_i(T) + \sum_{t=t_0+1}^{T}\{I_t = x_i, \rho_i(t)+s_i(t) \geq u_i(t)\}\\
&+ \sum_{t=t_0+1}^{T}\{A_t = x_i, \rho_i(t)+s_i(t) \geq u_i(t)\}\\
\leq& u_i(T) + \sum_{t=1}^{\infty}\{\mathcal{F}_t \vee \mathcal{G}_t\} + 1\\
&+ \underbrace{\sum_{t=t_0+2}^{T}\{A_t = x_i, \rho_i(t-1)+s_i(t-1) \geq u_i(t-1)\}}_{\Gamma}.
\end{aligned} \tag{10}$$

where t_0 ($1 \leq t_0 \leq T$) denotes the time when $\rho_i(t_0)+s_i(t_0) = u_i(T)$ holds. If such t_0 does not exist, the inequality still holds.

Event $A_t = x_i$ occurring implies that in MultiSBM-Feedback, the played pair of arms is (x_i, x_1). This occurs only if x_i was the right one in the previously played pair of arms, *i.e.*, (z_t, x_i). In other words, some SBM S_{z_t} has advanced suboptimal arm x_i before event $A_t = x_i$ occurs. However, at time-step $t = t_0+1, \ldots, T$, S_{x_1} will not advance x_i unless either of event $\mathcal{F}_t$ or $\mathcal{G}_t$ occurs. Thus, we can bound term Γ by

$$\begin{aligned}
\Gamma =& \sum_{t=t_0+2}^{T}\{A_t = x_i, \rho_i(t-1)+s_i(t-1) \geq u_i(t-1), z_t = x_1\}\\
&+ \sum_{t=t_0+2}^{T}\{A_t = x_i, \rho_i(t-1)+s_i(t-1) \geq u_i(t-1), z_t \neq x_1\}\\
\leq& \sum_{t=t_0+2}^{T}\{A_t = x_i, \rho_i(t-1)+s_i(t-1) \geq u_i(t-1), I_{t-1} = x_i\}\\
&+ \sum_{z\neq x_1}\sum_{t=1}^{\widetilde{T}}\{S_z \text{ advance } x_i\}
\end{aligned}$$

$$\leq \sum_{t=1}^{\infty}\{\mathcal{F}_t \vee \mathcal{G}_t\} + \sum_{z\neq x_1}\tau_{zx_i}(\tau_z(\widetilde{T})),$$

where $\widetilde{T}$ denotes the external time in MultiSBM-Feedback when the internal time in SBM S_{x_1} is T, *i.e.*, $\tau_{x_1}(\widetilde{T}) = T$.

Thus, Eq. (10) can be bounded by

$$\begin{aligned}&\rho_i(t) + s_i(t)\\ \leq& u_i(T) + 2\sum_{t=1}^{\infty}\{\mathcal{F}_t \vee \mathcal{G}_t\} + 1 + \sum_{z\neq x_1}\tau_{zx_i}(\tau_z(\widetilde{T})).\end{aligned}$$

Taking summation over $i > 1$ and using Eq. (9), we have

$$\begin{aligned}&\mathbb{E}[2\sum_{i>1}\rho_i(T)]\\ =&\mathbb{E}[\sum_{i>1}\rho_i(T) + \sum_{i>1}s_i(T)]\\ \leq&\sum_{i>1}u_i(T) + \frac{8}{\beta-2}K + K + \sum_{x_i\neq x_1}\sum_{z\neq x_1}\mathbb{E}[\tau_{zx_i}(\tau_z(\widetilde{T}))]\\ \leq&\sum_{i>1}\frac{2\beta}{\Delta_i^2}\ln T + \frac{\beta+6}{\beta-2}K + \sum_{x_i\neq x_1}\sum_{z\neq x_1}\mathbb{E}[\tau_{zx_i}(\tau_z(\widetilde{T}))].\end{aligned}$$

Since $\beta := \alpha + 2$, we have

$$\begin{aligned}\mathbb{E}[\sum_{i>1}\rho_i(T)] \leq& \sum_{i>1}\frac{(\alpha+2)}{\Delta_i^2}\ln T + \frac{\alpha+8}{2\alpha}K\\ &+ \frac{1}{2}\sum_{x_i\neq x_1}\sum_{z\neq x_1}\mathbb{E}[\tau_{zx_i}(\tau_z(\widetilde{T}))].\end{aligned}$$

Therefore, we can bound $\mathbb{E}[R_{x_1}(T)]$ by

$$\begin{aligned}\mathbb{E}[R_{x_1}(T)] \leq&\mathbb{E}[\sum_{i>1}\rho_i(T)]] \cdot \Delta_{max}\\ \leq& \sum_{i>1}\frac{(\alpha+2)\Delta_{max}}{\Delta_i^2}\ln T + \frac{(\alpha+8)\Delta_{max}}{2\alpha}K\\ &+ \frac{\Delta_{max}}{2}\sum_{y\neq x_1}\sum_{x\neq x_1}\mathbb{E}[\tau_{xy}(\tau_x(\widetilde{T}))].\end{aligned}$$

Using Eq. (8), the expected regret of MultiSBM-Feedback up to time $\widetilde{T}$ is bounded by

$$\begin{aligned}&\mathbb{E}[R(\widetilde{T})]\\ \overset{(a)}{\leq}&0.5 + \mathbb{E}[R_{x_1}(T)] + \mathbb{E}[\sum_{y\neq x_1}\sum_{x\neq x_1}R_{xy}(\tau_x(\widetilde{T}))]\\ \leq&0.5 + \sum_{i>1}\frac{(\alpha+2)\Delta_{max}}{\Delta_i^2}\ln T + \frac{(\alpha+8)\Delta_{max}}{2\alpha}K\\ &+ \frac{\Delta_{max}}{2}\sum_{y\neq x_1}\sum_{x\neq x_1}\mathbb{E}[\tau_{xy}(\tau_x(\widetilde{T}))]\\ &+ \sum_{y\neq x_1}\sum_{x\neq x_1}\mathbb{E}[R_{xy}(\tau_x(\widetilde{T}))]\\ \leq&\sum_{i>1}\frac{(\alpha+2)\Delta_{max}}{\Delta_i^2}\ln\widetilde{T} + \frac{(\alpha+8)\Delta_{max}}{2\alpha}K\\ &+ \sum_{j>1}\sum_{i>1}O\Big(\frac{\alpha\Delta_{max}}{\Delta_j^2}\big(\ln\ln\widetilde{T} + \ln K + \ln(\frac{1}{\Delta_i})\big)\Big),\end{aligned}$$

where (a) holds since $\tau_{x_1}(\widetilde{T}) = T$.

Replacing $\widetilde{T}$ with T, we obtain

$$\begin{aligned}\mathbb{E}[R(T)] \leq& \sum_{i>1}\frac{(\alpha+2)\Delta_{max}}{\Delta_i^2}\ln T + \frac{(\alpha+8)\Delta_{max}}{2\alpha}K\\ &+ \sum_{j>1}\sum_{i>1}O\Big(\frac{\alpha\Delta_{max}}{\Delta_j^2}\big(\ln\ln T + \ln K + \ln(\frac{1}{\Delta_i})\big)\Big). \quad (11)\end{aligned}$$

Note that our analysis uses a novel technique to bound $\rho_i(t) + s_i(t)$. In addition, the standard analysis procedure that bounds $\rho_i(t)$ by $u_i(t)$ in MultiSBM [3] still holds in our analysis. Thus, $\mathbb{E}[R_{x_1}(T)]$ can be also bounded by

$$\mathbb{E}[R_{x_1}(T)] \leq \sum_{i>1}\frac{2(\alpha+2)}{\Delta_i}\ln T + \frac{4\Delta_{max}}{\alpha}K.$$

Then, we can obtain

$$\begin{aligned}\mathbb{E}[R(T)] \leq&0.5 + \mathbb{E}[R_{x_1}(T)] + \mathbb{E}[\sum_{y\neq x_1}\sum_{x\neq x_1}R_{xy}(T)]\\ \leq& \sum_{i>1}\frac{2(\alpha+2)}{\Delta_i}\ln T + \frac{4\Delta_{max}}{\alpha}K\\ &+ \sum_{j>1}\sum_{i>1}O\Big(\frac{\alpha}{\Delta_y}\big(\ln\ln T + \ln K + \ln(\frac{1}{\Delta_x})\big)\Big). \quad (12)\end{aligned}$$

Theorem 3.3 follows from Eqs. (11) and (12).

□

C PROOF OF THEOREM 4.1

In order to prove Theorem 4.1, we quote a lemma (Lemma 7.2) in [30] and introduce another five lemmas (Lemmas 7.3 to 7.6 and 7.8) as follows.

Lemma 7.2. *Let $\mathbf{P} := [p_{ij}]$ be the preference matrix of a K-armed dueling bandit problem with arms $\{x_1, ..., x_K\}$. Then, for any dueling bandit algorithm and any $\alpha > \frac{1}{2}$ and $\delta > 0$, we have*

$$P\Big(\forall t > C(\delta), i, j, p_{ij} \in [l_{ij}(t), u_{ij}(t)]\Big) > 1 - \delta.$$

This lemma is quoted from (Lemma 1 in [30]).

Lemma 7.3. *With probability at least $1-\delta$, $\forall t > C(\delta)$, $x_1 \in C$.*

Proof. Using Lemma 7.2, we have that with probability at least $1-\delta$, $\forall t > C(\delta)$, i, $u_{1i}(t) \geq p_{1i}(t) \geq \frac{1}{2}$. This concludes the proof of Lemma 7.3. □

For ease of notation, we recall the notation defined in the proof sketch of Theorem 4.1. Case (a), (b) and (c) denote the three mutually exclusive cases corresponding to Line 11, Line 14 and Line 16 in Algorithm 4, respectively. Case (c-1) and (c-2) respectively denote the two mutually exclusive situations in case (c), *i.e.*, $x_1 \in \mathcal{A}_t$ and $x_1 \notin \mathcal{A}_t$.

We also introduce the following notation. Let $N(t)$ denote the number of times $\mathcal{A}_t \neq \{x_1\}$ up to time t. Let $\widetilde{N}_{ij}(t)$ ($\widetilde{N}_{ij}(t)$=$\widetilde{N}_{ji}(t)$) denote the number of observed dueling outcomes of x_i and x_j between time $C(\delta)+1$ and t. Let $\widetilde{N}^b(t)$, $\widetilde{N}_1^c(t)$ and $\widetilde{N}_2^c(t)$ respectively denote the number of times case (b), (c-1) and (c-2) occur between time $C(\delta)+1$ and t.

From Lemma 7.3, we know that with probability at least $1-\delta$, $\forall t > C(\delta)$, MultiRUCB will not carry out Line 8 in Algorithm 4 but one of case (a), case (b) and case (c). In addition, with probability at least $1-\delta$, $\forall t > C(\delta)$, if MultiRUCB carries out case (a), the comparison set will be $\mathcal{A}_t = \{x_1\}$. Thus, in order to bound the expected regret with probability at least $1-\delta$, it suffices to bound $\widetilde{N}^b(t)$, $\widetilde{N}_1^c(t)$ and $\widetilde{N}_2^c(t)$.

Lemma 7.4. *With probability at least* $1-\delta$, $\forall t > C(\delta)$,

$$\widetilde{N}^b(t) + \widetilde{N}_1^c(t) \leq \sum_{i>1} \frac{4\alpha}{\Delta_i^2} \ln t.$$

Proof. According to Lemma 7.3, with probability at least $1-\delta$, $\forall t > C(\delta)$, every time case (b) occurs, we can observe at least one outcome of duel between x_1 and some x_i $(i > 1)$ ($\sum_{i>1} \widetilde{N}_{1i}(t)$ increments by 1). Every time case (c-1) occurs, we can observe outcomes of $m-1$ duels between x_1 and x_i $(i > 1)$ ($\sum_{i>1} \widetilde{N}_{1i}(t)$ increments by $m-1$).

In the following we prove that with probability at least $1-\delta$, $\forall t > C(\delta), i > 1$, $\widetilde{N}_{1i}(t) \leq \frac{4\alpha}{\Delta_i^2} \ln t$.

Assume that $\exists t > C(\delta), \exists i > 1$. $\widetilde{N}_{1i}(t) > \frac{4\alpha}{\Delta_i^2} \ln t$. Let s denote the last time when we observed the dueling outcome between x_1 and x_i up to time t, which implies $\widetilde{N}_{1i}(s) = \widetilde{N}_{1i}(t), C(\delta) < s \leq t$. Using Lemma 7.2, we have that with probability at least $1-\delta$,

$$2\sqrt{\frac{\alpha \ln s}{N_{1i}(s)}} \leq 2\sqrt{\frac{\alpha \ln s}{\widetilde{N}_{1i}(s)}} \leq 2\sqrt{\frac{\alpha \ln t}{\widetilde{N}_{1i}(t)}} < \Delta_i$$

$$u_{i1}(s) \leq p_{i1} + 2\sqrt{\frac{\alpha \ln s}{N_{1i}(t)}} < p_{i1} + \Delta_i = \frac{1}{2}.$$

Since $u_{i1}(s) < \frac{1}{2}$, x_i cannot be in C and thus cannot be chosen into $\mathcal{A}_t$, which yields a contradiction.

Taking summation over $i > 1$, we have that with probability at least $1-\delta$, $\forall t > C(\delta)$, $\sum_{i>1} \widetilde{N}_{1i}(t) \leq \sum_{i>1} \frac{4\alpha}{\Delta_i^2} \ln t$.

Because with probability at least $1-\delta, \forall t > C(\delta)$, each occurrence of case (c-1) increments $\sum_{i>1} \widetilde{N}_{1i}(t)$ by $m-1$, we can bound the $\widetilde{N}_1^c(t)$ by

$$(m-1)\widetilde{N}_1^c(t) \leq \sum_{i>1} \widetilde{N}_{1i}(t) \leq \sum_{i>1} \frac{4\alpha}{\Delta_i^2} \ln t$$

$$\widetilde{N}_1^c(t) \leq \sum_{i>1} \frac{4\alpha}{(m-1)\Delta_i^2} \ln t. \tag{13}$$

Eq. (13) will be used later (proof of Lemma 7.8).

For $\widetilde{N}^b(t) + \widetilde{N}_1^c(t)$, we have that with probability at least $1-\delta$, $\forall t > C(\delta)$, $\widetilde{N}^b(t) + \widetilde{N}_1^c(t) \leq \sum_{i>1} \widetilde{N}_{1i}(t) \leq \sum_{i>1} \frac{4\alpha}{\Delta_i^2} \ln t$.

□

Lemma 7.5. *With probability at least* $1-\delta$, $\forall t > C(\delta)$,

$$\widetilde{N}_2^c(t) \leq \sum_{1<i<j} \frac{4\alpha}{C_m^2 \Delta_{ij}^2} \ln t.$$

Proof. $\forall t > C(\delta)$, every time case (c-2) occurs, we will observe outcomes of $C_m^2 := \frac{m(m-1)}{2}$ different duels between suboptimal arms, *i.e.*, $\sum_{1<i<j} \widetilde{N}_{ij}(t)$ will increment by C_m^2.

In the following we prove that with probability at least $1-\delta$, $\forall t > C(\delta), i, j > 1, i \neq j$, $\widetilde{N}_{ij}(t) \leq \frac{4\alpha}{\Delta_{ij}^2} \ln t$.

Assume that $\exists t > C(\delta), \exists i, j > 1,\ i \neq j$ (wlog $i < j$), $\widetilde{N}_{ij}(t) > \frac{4\alpha}{\Delta_{ij}^2} \ln t$. Let s denote the last time when we observed the dueling outcome between x_i and x_j up to time t, which implies $\widetilde{N}_{ij}(s) = \widetilde{N}_{ij}(t), C(\delta) < s \leq t$. Using Lemma 7.2, we have that with probability at least $1-\delta$,

$$2\sqrt{\frac{\alpha \ln s}{N_{ij}(s)}} \leq 2\sqrt{\frac{\alpha \ln s}{\widetilde{N}_{ij}(s)}} \leq 2\sqrt{\frac{\alpha \ln t}{\widetilde{N}_{ij}(t)}} < \Delta_{ij}$$

$$u_{ji}(s) \leq p_{ji} + 2\sqrt{\frac{\alpha \ln s}{N_{ij}(t)}} < p_{ji} + \Delta_{ij} = \frac{1}{2}.$$

Since $u_{ji}(s) < \frac{1}{2}$, x_j cannot be in C and thus cannot be chosen into $\mathcal{A}_t$, which yields a contradiction.

Taking summation over $1 < i < j$, we have that with probability at least $1-\delta$, $\forall t > C(\delta)$, $\sum_{1<i<j} \widetilde{N}_{ij}(t) \leq \sum_{1<i<j} \frac{4\alpha}{\Delta_{ij}^2} \ln t$. Thus, with probability at least $1-\delta$, $\forall t > C(\delta)$, $C_m^2 \widetilde{N}_2^c(t) \leq \sum_{1<i<j} \widetilde{N}_{ij}(t) \leq \sum_{1<i<j} \frac{4\alpha}{\Delta_{ij}^2} \ln t$. This concludes the proof of Lemma 7.5.

□

Lemma 7.6. *With probability at least* $1-\delta$, *for any time* T,

$$N(T) \leq C(\delta) + \sum_{i>1} \frac{4\alpha}{\Delta_i^2} \ln T + \sum_{1<i<j} \frac{4\alpha}{C_m^2 \Delta_{ij}^2} \ln T.$$

Proof. Lemma 7.6 holds by combining Lemma 7.4 and Lemma 7.5.

□

Lemma 7.6 gives a high probability bound of $N(T)$. In the following we will give another high probability bound (Lemma 7.8) of $N(T)$ using the choice strategy of case (c).

Before stating Lemma 7.8, we firstly introduce a definition, which will be used in the proof of Lemma 7.8.

Definition 7.7. Let $\widehat{T}_\delta$ be the smallest time satisfying

$$\widehat{T}_\delta > C(\frac{\delta}{2}) + \sum_{i>1} \frac{4\alpha}{\Delta_i^2} \ln \widehat{T}_\delta + \sum_{1<i<j} \frac{4\alpha}{C_m^2 \Delta_{ij}^2} \ln \widehat{T}_\delta.$$

where $\widehat{T}_\delta$ is guaranteed to exist because the left side of the inequality grows linearly with $\widehat{T}_\delta$ and the right side grows logarithmically.

In the following we prove a upper bound of $\widehat{T}_\delta$ using similar techniques in [30].

Define $C := C(\frac{\delta}{2})$, $D := \sum_{i>1} \frac{4\alpha}{\Delta_i^2} + \sum_{1<i<j} \frac{4\alpha}{C_m^2 \Delta_{ij}^2}$. To find a upper bound of $\widehat{T}_\delta$, we need to produce one number T satisfying $T > C + D \ln T$. It is easy to prove one such number is $T = 2C + 2D \ln 2D$.

$$C + D\ln(2C + 2D\ln 2D) \overset{a}{\leq} C + D\ln(2D\ln 2D) + D\frac{2C}{2D\ln 2D}$$

$$\overset{b}{\leq} C + D\ln((2D)^2) + \frac{C}{\ln 2D} \overset{c}{\leq} 2C + 2D\ln 2D,$$

where (a) uses a first order Taylor expansion, (b) uses $\ln 2D < 2D$ and (c) uses $D > 2$.

Thus, we can bound $\widehat{T}_\delta$ by

$$\widehat{T}_\delta \leq 2C + 2D\ln 2D. \tag{14}$$

Lemma 7.8. *With probability at least $1-\delta$, for any time T,*

$$N(T) \leq 2C + 2D\ln 2D + 4\ln\frac{2}{\delta} + \sum_{i>1}\frac{4\alpha}{\Delta_i^2}\ln T + 2\sum_{i>1}\frac{4\alpha}{(m-1)\Delta_i^2}\ln T.$$

where $C := C(\frac{\delta}{2})$, $D := \sum_{i>1}\frac{4\alpha}{\Delta_i^2} + \sum_{1<i<j}\frac{4\alpha}{C_m^2\Delta_{ij}^2}$.

Proof. From Lemmas 7.3 to 7.5 and Definition 7.7, we have that with probability at least $1-\frac{\delta}{2}$, there exists a time $T_\delta \in (C(\frac{\delta}{2}), \widehat{T}_\delta]$ when case (a) occurs. This implies that with probability at least $1-\frac{\delta}{2}$, $\mathcal{B}$ has been set as $\mathcal{B} = \{x_1\}$ from time T_δ on. Thus, from time T_δ on, if MultiRUCB carries out case (c), case(c-1) will occur with probability of $\frac{1}{2}$.

Let $\widehat{N}^b(t)$, $\widehat{N}_1^c(t)$ and $\widehat{N}_2^c(t)$ denote the number of times case (b), (c-1) and (c-2) occur between time $T_\delta+1$ and t, respectively. We also introduce the following two sets of random variables:

- $\tau_0, \tau_1, \tau_2, \ldots$, where $\tau_0 := T_\delta$ and τ_l is the l^{th} time case (c-1) occurs after time T_δ.
- $n_1, n_2, \ldots$, where n_l is the number of times case (c-2) occurs between τ_{l-1} and τ_l.

Using Lemma 7.4, we have that with probability at least $1-\frac{\delta}{2}$, $\forall t > T_\delta$,

$$\widehat{N}^b(t) + \widehat{N}_1^c(t) \leq \widetilde{N}^b(t) + \widetilde{N}_1^c(t) \leq \sum_{i>1}\frac{4\alpha}{\Delta_i^2}\ln t. \tag{15}$$

Using Eq. (13), we have that with probability at least $1-\frac{\delta}{2}$, $\forall t > T_\delta$,

$$\widehat{N}_1^c(t) \leq \widetilde{N}_1^c(t) \leq \sum_{i>1}\frac{4\alpha}{(m-1)\Delta_i^2}\ln t.$$

This means that with probability at least $1-\frac{\delta}{2}$, between time $T_\delta+1$ and t, case (c-1) occurs at most $L_1^c(t) := \sum_{i>1}\frac{4\alpha}{(m-1)\Delta_i^2}\ln t$ times. Moreover, with probability at least $1-\frac{\delta}{2}$, for any time $t > T_\delta$, if case (c-1) has occurred $L_1^c(t)$ times, all suboptimal arms x_i ($i>1$) satisfy $u_{i1} < \frac{1}{2}$ and case (c-2) cannot occur. Thus, we have that with probability at least $1-\frac{\delta}{2}$, $\forall t > T_\delta$,

$$\widehat{N}_2^c(t) \leq \sum_{l=1}^{L_1^c(t)} n_l.$$

To bound the sum of intervals n_l, we introduce i.i.d. geometric random variables $\{\hat{n}_l\}_{l=1,2,\ldots,r}$ with parameter $\frac{1}{2}$. $\hat{n}_l$ bounds n_l because n_l counts the number of times it takes for case (c) to produce one case (c-1).

Using similar techniques in [10, 30] to bound the sum of $\{\hat{n}_l\}_{l=1,2,\ldots,r}$, which we denote by n, we can obtain that with probability at least $1-\frac{\delta}{2}$,

$$n < 2r + 4\ln\frac{2}{\delta},$$

Note that this $1-\frac{\delta}{2}$ is different from the aforementioned $1-\frac{\delta}{2}$ that is derived from Lemma 7.2.

Setting $r = L_1^c(t)$, we have that with probability at least $1-\delta$, $\forall t > T_\delta$,

$$\widehat{N}_2^c(t) \leq \sum_{l=1}^{L_1^c(t)} n_l \leq 2\sum_{i>1}\frac{4\alpha}{(m-1)\Delta_i^2}\ln t + 4\ln\frac{2}{\delta}. \tag{16}$$

Taking summation over T_δ (Eq. (14)), $\widehat{N}^b(t)$, $\widehat{N}_1^c(t)$ (Eq. (15)) and $\widehat{N}_2^c(t)$ (Eq. (16)), we obtain the result of Lemma 7.8.

□

Theorem 4.1. *Consider a K-armed multi-dueling bandits game, where the number of comparing arms is at most m at every time. Given $\alpha > 1$, the expected regret of MultiRUCB is bounded by*

$$\mathbb{E}[R_T] \leq \left[\left(\frac{2(4\alpha-1)K^2}{2\alpha-1}\right)^{\frac{1}{2\alpha-1}}\frac{2\alpha-1}{\alpha-1}\right]\Delta_{max} + \min\left\{D\Delta_{max}\ln T, \left(8 + 2D\ln 2D\right)\Delta_{max} + \frac{m+1}{m-1}\sum_{i>1}\frac{4\alpha\Delta_{max}}{\Delta_i^2}\ln T\right\},$$

where $D := \sum_{i>1}\frac{4\alpha}{\Delta_i^2} + \sum_{1<i<j}\frac{4\alpha}{C_m^2\Delta_{ij}^2}$ and $C_m^2 := \frac{m(m-1)}{2}$.

Proof. Combining Lemma 7.6 and Lemma 7.8, we can obtain that with probability at least $1-\delta$, for any time T,

$$N(T) \leq \min\left\{C(\delta) + D\ln T,\ 2C(\frac{\delta}{2}) + 2D\ln 2D + 4\ln\frac{2}{\delta} + \sum_{i>1}\frac{4\alpha}{\Delta_i^2}\ln T + 2\sum_{i>1}\frac{4\alpha}{(m-1)\Delta_i^2}\ln T\right\},$$

Integrating $N(T)$ with respect to δ from 0 to 1, we have that given $\alpha > 1$, $\mathbb{E}[N(T)]$ is bounded by

$$\mathbb{E}[N(T)] \leq \left[\left(\frac{2(4\alpha-1)K^2}{2\alpha-1}\right)^{\frac{1}{2\alpha-1}}\frac{2\alpha-1}{\alpha-1}\right] + \min\left\{D\ln T, 2D\ln 2D + 8 + \frac{m+1}{m-1}\sum_{i>1}\frac{4\alpha}{\Delta_i^2}\ln T\right\}.$$

Theorem 4.1 is obtained by applying

$$\mathbb{E}[R_T] \leq \mathbb{E}[N(T)]\cdot\Delta_{max}.$$

□

D VARIANCE RESULTS IN MULTI-DUELING BANDITS EXPERIMENTS

In this section, we present the omitted variance results in the multi-dueling bandits experiments (See Section 6.2). Table 1 shows the variances of cumulative regrets at the 10^6 timestep for 50 independent runs. Columns 2-5 correspond to the experiments in Figure 2 (a-d). "Syn" refers to the synthetic dataset.

Table 1: The variance results in the multi-dueling bandit experiments.

Algorithms	Syn, m=8	Syn, m=16	MSLR, m=8	MSLR, m=16
MultiRUCB	1648.05	1042.07	678.42	231.97
IndSelfSparring	2047.69	1243.43	741.39	264.67
MDB	2142.22	1246.30	754.21	277.62
MultiSparring	2201.16	1328.71	864.56	392.61