Supplementary: Battle of Bandits

A Proof of Lemma 1

Proof. We start by noting that

$$\begin{aligned}
&\mathbf{P}(X=i) = \mathbf{P}(X=i, i\in\{U,V\}) \\
&= \mathbf{P}(i\in\{U,V\})\,\mathbf{P}(X=i|i\in\{U,V\}) \\
&= \frac{(k-1)}{\binom{k}{2}}\,\mathbf{P}(X=i|i\in\{U,V\}) \\
&= \frac{(k-1)}{\binom{k}{2}}\sum_{j=1,j\neq i}^{k}\mathbf{P}(X=i,\{U,V\}=\{i,j\}|i\in\{U,V\}) \\
&= \frac{(k-1)}{\binom{k}{2}}\sum_{j=1,j\neq i}^{k}\mathbf{P}(\{U,V\}=\{i,j\}|i\in\{U,V\})\times \\
&\qquad\qquad \mathbf{P}(X=i|\{U,V\}=\{i,j\}) \\
&= \frac{(k-1)}{\binom{k}{2}}\sum_{j=1,j\neq i}^{k}\frac{1}{(k-1)}\times Q_{a_ia_j} = \sum_{j=1,j\neq i}^{k}\frac{2Q_{a_ia_j}}{k(k-1)},
\end{aligned}$$

where the penultimate equality is by appealing to sampling without replacement and the definition of the win probability for i in the pairwise preference model. □

B Regret analysis of Battling-Doubler

We will be using the following regret guarantee of UCB algorithm for classical MAB problem in order to proof the regret bounds of *Battling-Doubler*:

Theorem 12. ***UCB Regret [5].*** *Assume $[n]$ denotes the set of arms. μ_i denotes the expected reward associated with arm $i\in[n]$, such that $\mu_1>\mu_2\geq\cdots\geq\mu_n$ and $H=\sum_{i=2}^{n}\frac{1}{\Delta_i}$, where $\Delta_i=\mu_1-\mu_i$, then the expected regret of the UCB algorithm with confidence parameter $\zeta>0$ is of $O(\zeta H\ln T)$.*

Proof of Theorem 3

Proof. For any time horizon $T\in\mathbb{Z}_+$, let $B(T)$ denotes the supremum of the expected regret of the SBM $\mathcal{S}$ after T steps, over all possible reward distributions on the arm set $[n]$. Clearly, length of epoch ℓ in the algorithm is exactly $2^\ell,\ \forall\ell$. We denote this by $T_\ell=2^\ell$. Note that for all time steps t inside this epoch, the first $k-1$ arms $a_1^t,a_2^t,\cdots,a_{k-1}^t$ are chosen uniformly from $\mathcal{L}$, and thus the observed reward of the k^{th} arm also comes from a fixed distribution, as argued below:

Let $\theta'=\mathbf{E}_{a\sim Unif(\mathcal{L})}[\theta_a]$ denote the expected utility score value for each of the first $k-1$ arms in all steps t within epoch ℓ. Clearly, $\mathbf{E}[\theta_{a_1^t}]=\mathbf{E}[\theta_{a_2^t}]=\cdots=\mathbf{E}[\theta_{a_{k-1}^t}]=\theta'$, since each of the first $(k-1)$ arms were drawn uniformly from $\mathcal{L}$. Now, the SBM $\mathcal{S}$ is playing a standard MAB game over the set of arms $[n]$ with binary rewards. Let b_t denote the binary reward of the k^{th} arm a_k^t in the t^{th} step. Clearly for all t within epoch ℓ,

$$\mathbf{E}[b_t|(a_1^t,a_2^t,\cdots,a_k^t)]=\frac{1}{k}+\frac{\sum_{j=1}^{k-1}(\theta_{a_k^t}-\theta_{a_j^t})}{k(k-1)}$$

Thus at round t, if $a_k^t=x$, for any $x\in[n]$,

$$\begin{aligned}
\mathbf{E}[b_t|a_k^t=x] &= \mathbf{E}\left[\frac{1}{k}+\frac{\sum_{j=1}^{k-1}(\theta_{a_x}-\theta_{a_j^t})}{k(k-1)}\right] \\
&= \left[\frac{1}{k}+\frac{(k-1)(\theta_{a_x}-\theta')}{k(k-1)}\right]=\left[\frac{1}{k}+\frac{(\theta_{a_x}-\theta')}{k}\right].
\end{aligned}$$

Now for SBM $\mathcal{S}$, the best arm with highest expected reward is still arm-1, where

$$\mathbf{E}[b_t|a_k^t=1]=\left[\frac{1}{k}+\frac{(\theta_1-\theta')}{k}\right] \tag{5}$$

By the definition of the bound function B(T), the total expected regret (in the traditional MAB sense) of the SBM $\mathcal{S}$, in the epoch ℓ is at most $B(T_\ell)=B(2^\ell)$, which implies that

$$\mathbf{E}_x\left[\left(\frac{1}{k}+\frac{(\theta_1-\theta')}{k}\right)-\left(\frac{1}{k}+\frac{(\theta_{a_x}-\theta')}{k}\right)\right]\leq B(2^\ell)$$

Thus,

$$\mathbf{E}_x\left[\left(\frac{\theta_1-\theta_{a_x}}{k}\right)\right]\leq B(2^\ell)$$

This in other words says that the expected contribution of the k^{th} arm to the regret in phase i is at most $B(2^\ell)$. It only remains to bound the expected contribution of the remaining $(k-1)$ arms to the regret, which are drawn from a distribution that assigns to all arms $a\in[n]$ a probability proportional to the frequency in which a is played as the k^{th} arm in the previous epoch $(\ell-1)$.

The interesting thing to note is that, at any round t of epoch ℓ, the expected regret incurred by any of the first $k-1$ arms, i.e. a_j^t such that $j\in[k-1]$, is exactly same as the average expected regret contributed by the set of arms drawn as the k^{th} arm in epoch $\ell-1$ (since Line 7 selects $a_1^t,a_2^t,\cdots,a_{k-1}^t$ uniformly from $\mathcal{L}$ at any epoch ℓ of length 2^ℓ), and thus it is at most $B(2^{\ell-1})/2^{\ell-1}$.

Hence the total expected regret incurred by any of the first $k-1$ arms, $a_j^t,\ j \in [k-1]$ in epoch ℓ is bounded by $2^\ell(B(2^{\ell-1})/2^{\ell-1}) = 2B(2^{\ell-1})$. Since any finite time horizon of T can be uniquely decomposed as $2+4+8+\cdots+2^s+Z$, for some integer $s \geq 1$ and $0 \leq Z \leq 2^{s+1}-1$. Thus the total expected regret of Battling-Doubler is given by the following function of T:

$$\frac{k-1}{k} + 3(k-1)\big(B(2) + B(2^2) + \cdots B(2^s) + B(Z)\big) \tag{6}$$

Now by our assumption, $B(t) = O(\ln t)^\beta$, for any $t \in \mathbb{Z}_+$, the theorem claim follows by simplifying and bounding the above expression. □

Proof of corollary 4 Note that if the SBM $\mathcal{S}$ in Battling-Doubler is UCB that operates on the set of n arms, with expected reward of arm $i \in [n]$ being $\left(\frac{1}{k} + \frac{(\theta_1 - \theta')}{k}\right)$ (from equation 5). Thus for SBM $\mathcal{S}$, the complexity of the total gap of the suboptimal arms become $\sum_{i=2}^n \frac{1}{\Delta_i} = kH$. Now from Theorem 12, we have that $B(t) = O(\zeta kH \ln t)$ for any $t \in \mathbb{Z}_+$. Thus we get

$$\begin{aligned}
\mathbf{E}[R_T] &= \mathbf{E}\left[\sum_{t=1}^T \left(\frac{\sum_{j\in S_t}(\theta_1 - \theta_j)}{2k}\right)\right] \\
&\leq \frac{1}{2k}\Bigg(\frac{k-1}{k} + 3(k-1)\big(B(2) + B(2^2) + \\
&\qquad \cdots B(2^s) + B(Z)\big)\Bigg) \\
&\leq \frac{1}{2k}\Big(O(\zeta k^2 H \ln^2 T)\Big) = \Big(O(\zeta kH \ln^2 T)\Big).
\end{aligned}$$

Proof of Corollary 5 The claim simply follows from the above theorem and the fact that the worst possible gap for UCB algorithm can be at most $\Delta = O(\sqrt{\frac{n \ln T}{T}})$, as argued in [18] or [14] (see Corollary 1.1). The claim now follows by substituting $H = (n-1)\frac{1}{\Delta} = O(\sqrt{\frac{nT}{\ln T}})$ in Corollary 4.

C Regret analysis of Battling-MultiSBM

Proof of Theorem 6.

Proof. We start by noting that in Battling-MultiSBM, only one SBM advances at each round. Denote by $\rho_x(t)$ the total number of times $\mathcal{S}_x$ was advanced after t iterations of the algorithm, for any arm $x \in [n]$.

Then at any round t where x is played as the $(k-1)^{th}$ arm, $\mathcal{S}_x$ internally sees a world in which the rewards are binary, and the expected reward of any arm $a \in [n]$ is exactly $\frac{1}{k} + \frac{\sum_{j=1}^{k-1}\left(\theta_a - \theta_{a_j^t}\right)}{k(k-1)} = \frac{1}{k} + \frac{\sum_{j=t-k+1}^{t-1}\left(\theta_a - \theta_{a_k^j}\right)}{k(k-1)}$ (from (2) and Line 5 of Battling-MultiSBM). Note that for all SBMs $\mathcal{S}_y,\ y \in [n]$, the best arm (the one with highest expected reward) to play is arm 1. Thus at round t, the reward corresponding to the best arm is $\frac{1}{k} + \frac{\sum_{j=t-k+1}^{t-1}\left(\theta_1 - \theta_{a_k^j}\right)}{k(k-1)}$.

Now it is easy to see that for all SBMs $\mathcal{S}(y),\ y \in [n]$, the suboptimalities of the arms are the same: the suboptimality (regret) associated to a is $\frac{\left(\theta_1 - \theta_a\right)}{k}$, for all arm $a \in [n]$.

Following is the key observation in this proof, the total regret incurred by Battling-MultiSBM can be written as:

$$\begin{aligned}
R_T &= \sum_{t=1}^T \left(\frac{\sum_{j\in S_t}(\theta_1 - \theta_j)}{k}\right) \\
&= \sum_{t=1}^T \left(\frac{\sum_{j=1}^k (\theta_1 - \theta_{a_j^t})}{k}\right) \\
&= \frac{(\theta_1 - \theta_{a_2^0})}{k} + 2\frac{(\theta_1 - \theta_{a_3^0})}{k} + \cdots \\
&+ (k-1)\frac{(\theta_1 - \theta_{a_k^0})}{k} + k\sum_{t=1}^{T-k+1}\frac{(\theta_1 - \theta_{a_k^t})}{k} \\
&+ (k-1)\frac{(\theta_1 - \theta_{a_k^{T-k+2}})}{k} + (k-2)\frac{(\theta_1 - \theta_{a_k^{T-k+3}})}{k} \\
&\quad + \cdots + \frac{(\theta_1 - \theta_{a_k^T})}{k} \\
&\leq k\sum_{t=1}^T \frac{(\theta_1 - \theta_{a_k^t})}{k} + (k-1) \qquad (7) \\
&\qquad \left(\text{as, } \max_{a\in[n]}(\theta_1 - \theta_a) \leq 1\right)
\end{aligned}$$

This essentially conveys that bounding the above regret is equivalent to bounding regret of each SBM $\mathcal{S}_y,\ y \in [k]$. In other words, this equivalently says that any suboptimal SBM $\mathcal{S}_y,\ y \in [n] \setminus \{1\}$ can not be played to many times, and any SBM can not play a suboptimal arm $a \in [n] \setminus \{1\}$ too many times. The rest of the proof justifies the above claim.

Let us denote by ρ_x, the total number of times SBM $\mathcal{S}_x$ is called, clearly $\rho_x = \sum_{t=1}^{T-1} \mathbf{1}(a_{k-1}^t \text{ outputs } x)$ and by ρ_{xy} the total number of times SBM $\mathcal{S}_x$ played arm y, i.e. $\rho_{xy} = \sum_{t=1}^{T-1} \mathbf{1}(a_{k-1}^t = x \text{ and } a_k^t = y)$.

We also denote by $R_x(T') = \frac{1}{k}\sum_{\{t \mid \rho_x \leq T',\ a_{k-1}^t = x\}}\left(\theta_1 - \theta_x\right)$, the regret incurred due to advancing SBM $\mathcal{S}_x$, till time T'. In words, this is the contribution of the k^{th} bandit choices to the regret at all times t for which the $(k-1)^{th}$ arm is chosen to be x, and $\mathcal{S}_x$'s internal

counter has not surpassed T'. Similarly we denote by $R_{xy}(T') = \frac{1}{k}\sum_{\{t \mid \rho_x \le T',\, a^t_{k-1}=x,\, a^t_k=y\}} \left(\theta_1 - \theta_y\right)$ the regret incurred due to SBM $\mathcal{S}_x$ for playing the arm $y \in [n]$, upto time T'. Clearly $R_{xy}(T') = \rho_{xy}(T')\frac{\Delta_y}{k}$.

From equation 7 it is now easy to see that, in order to bound the expected regret R_T, it suffices to bound the expressions $\mathbf{E}[R_{xy}(\rho_x(T))],\ \forall x, y \in [n]$.

Note that here we assume that each SBM policy used in Battling-MultiSBM is α-robust which would be crucially used in the proof. The main insight is to exploit the fact that due to α-robustness of the SBMs used, $\rho_x(T)$ is order of $\ln T$ for any suboptimal x. We begin with the observation that for any fixed $x, y \in [n]$, (x being a suboptimal arm), if we choose $\alpha > 2$, and $s \ge 4\alpha$, from the α-robustness property of the SBM we get,

$$\begin{aligned}
\mathbf{P}\left[R_{xy}(T) \;\ge\; \frac{(s\ln T)}{\Delta_y}\right] &= \mathbf{P}\left[R_{xy}(T) \ge \frac{((s/k)\ln T)}{(\Delta_y)/k}\right] \\
&= \mathbf{P}\left[\rho_{xy} \ge \frac{((s/k)\ln T)}{((\Delta_y)/k)^2}\right] \\
&\le \frac{2}{\alpha}\left[\frac{((s/k)\ln T)}{((\Delta_y)/k)^2}\right]^{-\alpha} \\
&= \frac{2}{\alpha}\left[\frac{((sk)\ln T)}{((\Delta_y))^2}\right]^{-\alpha} \\
&\le (sk\ln T)^{-\alpha}, \qquad (8)
\end{aligned}$$

where the last inequality follows due to the fact that $\Delta_y \le 1,\ \forall y \in [n]$ and $\alpha \ge 2$. Then under the same assumption on s and x, y, using union bound we get,

$$\mathbf{P}\left[\exists p \in \{0, \ldots, \lceil \ln\ln T\rceil\} : R_{xy}(e^{e^p}) \ge sp/\Delta_y\right] \le 2s^{-\alpha}$$

We now bound the quantity $\rho_x(T)$ for any non-optimal fixed x using the fact that all $z \in [n]$ satisfy $\rho_z(T) \le T$, and any SBM $\mathcal{S}_x$ is advanced in an iteration only if x was the k^{th} bandit arm in the previous round. Thus we have that for all suboptimal $x \in [n] \setminus \{1\}$,

$$\begin{aligned}
\mathbf{P}[\rho_x(T) \;\ge\; (sn\ln T)/\Delta_x^2] &= \mathbf{P}\left[\sum_{z\in[n]} \rho_{zx}(T) \ge (sn\ln T)/\Delta_x^2\right] \\
&\le \sum_{z\in[n]} \mathbf{P}\left[R_{zx}(T) \ge \frac{\left(\frac{s}{k}\ln T\right)}{\Delta_x}\right] \\
&\le n(s\ln T)^{-\alpha}, \qquad (9)
\end{aligned}$$

where the rightmost inequality is by union bound and (8). Now fix some $x, y \in [n]$, such that is x suboptimal. The last two inequalities give rise to a random variable Z defined as the minimal scalar for which we have for all $T' \in [e, e^e, e^{e^2}, \cdots, e^{e^{\lceil \ln\ln T\rceil}}]$,

$$\rho_x(T) \le \frac{Zn\ln T}{\Delta_x^2}, \text{ and } R_{xy}(T') \le \frac{Z\ln T'}{\Delta_y}.$$

By (8) and (9), we have that for all $s \ge 4k\alpha$, $\mathbf{P}[Z \ge s] \le 2s^{-\alpha} + n(s\ln T)^{-\alpha}$. Also, conditioned on the event that $\{Z \le s\}$, we have $R_{xy}(\rho_x(T)) \le R^s_{xy} := \frac{se\ln((sn\ln T)/\Delta_x^2)}{\Delta_y} = (se\Delta_y^{-1}(\ln\ln T + \ln n + \ln s - 2\ln\Delta_x))$. Combining above we get,

$$\begin{aligned}
&\mathbf{E}[R_{xy}(\rho_x(T))] \\
&= O(R_{xy}^{8\alpha-1} + \sum_{i=1}^{\infty} R_{xy}^{8\alpha+i} \\
&\qquad (2(4k\alpha+i)^{-\alpha} + n((4k\alpha+i)\ln T)^{-\alpha}).
\end{aligned}$$

Now since $\alpha = \max\{3, 2 + \frac{\ln n}{\ln\ln T}\}$, it can be verified that the last expression converges to $O(R_{xy}^{8\alpha})$, hence

$$\mathbf{E}[R_{xy}(\rho_x(T))] = O(\alpha\Delta_y^{-1}(\ln\ln T + \ln n - 2\ln\Delta_x)). \qquad (10)$$

Combining above and using (7) we get, the total expected regret $\mathbf{E}[R_T]$ is at most $\big(k - 1 + k\mathbf{E}[R_1(\rho_1(T)) + \sum_{x,y\in[n]\setminus\{1\}} R_{xy}(\rho_x(T))]\big)$. The desired regret bound now follows from (10). □

Proof of Corollary 7 Similar to the case of *Battling-Doubler* (Corollary 5), the current claim simply follows from the regret guarantee of *Battling-MultiSBM* and the fact that the worst case gap can be at most $O(\sqrt{\frac{n\ln T}{T}})$, as argued in [18] or [14] (see Corollary 1.1). The claim now follows by substituting $H = (n-1)\frac{1}{\Delta} = O(\sqrt{\frac{nT}{\ln T}})$ in Theorem 6.

D Regret analysis of Battling-Duel

Proof of Theorem 8

Proof. Without loss of generality we will assume that the Condorcet arm $a^* = 1$ throughout the proof.

Our goal is to analyse the regret of Battling-Duel in terms of its underlying dueling bandit algorithm. Considering $\mathbf{Q}'$ to be the pairwise preference matrix perceived by dueling bandit algorithm $\mathcal{D}$, i.e. upon playing any pair of item $(x_t, y_t) \in [n] \times [n]$, it receives feedback according to the pairwise preference $P(x_t \text{ beats } y_t) = Q'_{x_t,y_t}$, we

know that the regret seen by the underlying dueling bandit algorithm is given by $R_T^{DB} = \sum_{t=1}^{T} \frac{Q'_{1,x_t} + Q'_{1,y_t} - 1}{2}$. We now start by analysing the relation between $\mathbf{Q}'$ and $\mathbf{Q}$.

Case 1. k **is even.** This is the easy case, since note that at any round t both x_t and y_t are replicated exactly for $\frac{k}{2}$ times. Thus at any round t:

$$\begin{aligned} Q'(x_t, y_t) &= \sum_{i=1}^{k/2} P(i|S_t) \\ &= \left(\frac{k}{2}\right) 2\frac{\left(\frac{k}{2}-1\right) Q_{x_t,x_t} + \left(\frac{k}{2}\right) Q_{x_t,y_t}}{k(k-1)} \\ &= \frac{k}{2(k-1)} Q_{x_t,y_t} + \frac{1}{4}\frac{(k-2)}{(k-1)}, \end{aligned}$$

where the second equality follows from the definition of *pairwise-subset choice model* (see Lemma 1 for details) and the last equality follows from the fact that $Q_{i,i} = \frac{1}{2},\ \forall i \in [n]$. Similarly we get,

$$\begin{aligned} Q'(y_t, x_t) &= \sum_{l=\frac{k}{2}+1}^{k} P(l|S_t) \\ &= \left(\frac{k}{2}\right) 2\frac{\left(\frac{k}{2}-1\right) Q_{y_t,y_t} + \left(\frac{k}{2}\right) Q_{y_t,x_t}}{k(k-1)} \\ &= \frac{k}{2(k-1)} Q_{y_t,x_t} + \frac{1}{4}\frac{(k-2)}{(k-1)}. \end{aligned}$$

Note the above expressions hold true for any pair of items $(x_t, y_t) \in [n] \times [n]$. Also, its also worth noting that indeed these give $Q'_{i,j} + Q'_{j,i} = 1$, and $Q'_{i,i} = \frac{1}{2},\ \forall i, j \in [n]$. Thus for any pair of items (i, j), we have

$$Q'_{i,j} = \frac{k}{2(k-1)} Q_{i,j} + \frac{1}{4}\frac{(k-2)}{(k-1)}. \tag{11}$$

Now let us analyse the instantaneous regret of Battling-Duel at round t, $r_t^{BB}(BD)$; using our definition of regret as defined in (3), gives:

$$\begin{aligned} r_t^{BB}(BD) &= \frac{1}{k}\sum_{j \in S_t} \left(Q_{1,j} - \frac{1}{2}\right) \\ &= \frac{1}{k}\sum_{j \in S_t} (Q_{1,j}) - \frac{1}{2} \\ &= \frac{1}{k}\left(\frac{k}{2}\left(Q_{1,x_t} + Q_{1,x_t}\right)\right) - \frac{1}{2} \\ &= \frac{2(k-1)}{k}\left(\frac{\left(Q'_{1,x_t} - \frac{1}{2}\right) + \left(Q'_{1,y_t} - \frac{1}{2}\right)}{2}\right) \\ &= \frac{2(k-1)}{k} r_t^{DB}(\mathcal{D}), \end{aligned}$$

where the second last equality follows from Equation 11. Thus summing over $t = 1, 2, \ldots T$, the cumulative regret of *Battling-Duel (BD)* over T rounds become:

$$\begin{aligned} R_T^{BB}(BD) &= \sum_{t=1}^{T} r_t^{BB}(BD) \\ &= \frac{2(k-1)}{k}\sum_{t=1}^{T}\left(r_t^{DB}(\mathcal{D})\right) \\ &= \frac{2(k-1)}{k} R_T^{DB}(\mathcal{D}), \end{aligned}$$

and the claim follows. Now let us consider the case when k is odd.

Case 2. k **is odd.**

Note that, similar to the case before, we again have that at any round t,

$$\begin{aligned} Q'(x_t, y_t) &= \frac{1}{2}\sum_{i=1}^{(k-1)/2} P(i|S_t) + \frac{1}{2}\sum_{i=1}^{(k+1)/2} P(i|S_t) \\ &= \left(\frac{k-1}{2}\right)\left(\frac{1}{2}\frac{\left(\frac{k-1}{2}-1\right) Q_{x_t,x_t} + \left(\frac{k+1}{2}\right) Q_{x_t,y_t}}{k(k-1)/2}\right) \\ &\quad + \left(\frac{k+1}{2}\right)\left(\frac{1}{2}\frac{\left(\frac{k+1}{2}-1\right) Q_{x_t,x_t} + \left(\frac{k-1}{2}\right) Q_{x_t,y_t}}{k(k-1)/2}\right) \\ &= \frac{k+1}{2k} Q_{x_t,y_t} + \frac{k-1}{4k} \end{aligned}$$

Similarly one can show that $Q'_{y_t,x_t} = \frac{k+1}{2k} Q_{y_t,x_t} + \frac{k-1}{4k}$. It is easy to verify that as desired $Q'_{x_t,y_t} + Q'_{y_t,x_t} = 1$. Furthermore above relation also gives that

$$Q_{x_t,y_t} - \frac{1}{2} = \frac{2k}{k+1}\left(Q'_{x_t,y_t} - \frac{1}{2}\right). \tag{12}$$

Then similar to Case 1, analysing the instantaneous regret of Battling-Duel at round t, $r_t^{BB}(BD)$; using our definition of regret as defined in (3), gives:

$$\begin{aligned} r_t^{BB}(BD) &= \frac{1}{k}\sum_{j \in S_t} \left(Q_{1,j} - \frac{1}{2}\right) \\ &= \frac{1}{k}\sum_{j \in S_t} (Q_{1,j}) - \frac{1}{2} \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{k}\left(\frac{1}{2}\left(\frac{k-1}{2}Q_{1,x_t}+\frac{k+1}{2}Q_{1,y_t}\right)\right.\\
&+ \left.\frac{1}{2}\left(\frac{k+1}{2}Q_{1,x_t}+\frac{k-1}{2}Q_{1,y_t}\right)\right)-\frac{1}{2}\\
&= \frac{\left(Q_{1,x_t}-\frac{1}{2}\right)+\left(Q_{1,y_t}-\frac{1}{2}\right)}{2}\\
&= \frac{2k}{k+1}\frac{\left(Q'_{1,x_t}-\frac{1}{2}\right)+\left(Q'_{1,y_t}-\frac{1}{2}\right)}{2}\\
&= \frac{2k}{k+1}r_t^{DB}(\mathcal{D}),
\end{aligned}
$$

where the second last equality follows from Equation 12. Now summing over $t = 1, 2, \ldots T$, the cumulative regret of *Battling-Duel (BD)* over T rounds become:

$$
\begin{aligned}
R_T^{BB}(BD) &= \sum_{t=1}^{T} r_t^{BB}(BD)\\
&= \frac{2k}{k+1}\sum_{t=1}^{T}\left(r_t^{DB}(\mathcal{D})\right)\\
&= \frac{2k}{k+1}R_T^{DB}(\mathcal{D}),
\end{aligned}
$$

and the claim for Case 2 follows.

□

Proof of Corollary 9

Proof. The result is immediate from Theorem 8 along with the expected regret guarantee of the RUCB algorithm as derived in Theorem 5 of [24]. □

Proof of Corollary 11

Proof. The proof immediately follows from Theorem 10 along with the regret lower bound for any DB problem as derived in Theorem 2 of [13]. This is because any smaller regret for $\mathcal{A}_{BB}$ would violate the best achievable regret for DB, which is a logical contradiction. □

E Datasets for experiments

E.1 Parameters for *linear-subset choice model*

For synthetic experiments with *linear-subset choice model*, we use the following four different utility score vectors $\boldsymbol{\theta} \in [0, 1]^n$: 1. *arith* 2. *geom* 3. *g1* and 4. *g3*.

Both *arith* and *geom* utility score vector has $n = 8$ items, with item 1 as the 'best' (Condorcet) item with highest score, i.e. $\theta_1 > \max_{i=2}^{8}\theta_i$ and rest of the θ_is are in an arithmetic or geometric progression respectively, as their name suggests. The two score vectors are described in Table 2.

arith	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
geom	0.8	0.7	0.512	0.374	0.274	0.2	0.147	0.108

Table 2: Parameters for *linear-subset choice model*

The next two utility score vectors has $n = 15$ items in each. Similarly as before, item 1 is the Condorcet winner here as well, with $\theta_1 > \max_{i=2}^{8}\theta_i$. More specifically for *g1*, the individual score vectors are of the form:

$$
\theta_i = \begin{cases} 0.8, & \text{if } i = 1\\ 0.2, & \forall i \in [15] \setminus \{1\} \end{cases}
$$

For *g3* the individual score vectors are of the form:

$$
\theta_i = \begin{cases} 0.8, & \text{if } i = 1\\ 0.7, & \forall i \in [8] \setminus \{1\}\\ 0.6, & \text{otherwise} \end{cases}
$$

We also used a bigger version of the *g1* dataset with $n = 50$ items to run experiments with varying subset size k as shown in Figure 5 (Section 5.3). The individual score vectors of *g3* are the form:

$$
\theta_i = \begin{cases} 0.8, & \text{if } i = 1\\ 0.2, & \forall i \in [50] \setminus \{1\} \end{cases}
$$

E.2 Pairwise preference matrices used in synthetic experiments with *pairwise-subset choice model*

We run experiments on two synthetic preference matrices *arith-pref* and *arxiv-pref* for *pairwise-subset choice model*. The datasets are shown in Table 3 and 4 respectively.

Arith preference dataset

0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85
0.45	0.5	0.55	0.6	0.65	0.7	0.75	0.8
0.4	0.45	0.5	0.55	0.6	0.65	0.7	0.75
0.35	0.4	0.45	0.5	0.55	0.6	0.65	0.7
0.3	0.35	0.4	0.45	0.5	0.55	0.6	0.65
0.25	0.3	0.35	0.4	0.45	0.5	0.55	0.6
0.2	0.25	0.3	0.35	0.4	0.45	0.5	0.55
0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5

Table 3: arith-pref: Arith preference matrix

Arxiv preference dataset

0.5	0.55	0.55	0.54	0.61	0.61
0.45	0.5	0.55	0.55	0.58	0.6
0.45	0.45	0.5	0.54	0.51	0.56
0.46	0.45	0.46	0.5	0.54	0.5
0.39	0.42	0.49	0.46	0.5	0.51
0.39	0.4	0.44	0.5	0.49	0.5

Table 4: arxiv-pref: Arxiv preference matrix

E.3 Pairwise preference matrices used in real world experiments with *pairwise-subset choice model*

Hurdy dataset

The Hudry tournament data is a well-studied tournament on 13 items and which actually has a special property of being the smallest tournament dataset for which the Banks and Copeland sets are different as shown in [16]. Although the original dataset does not contain a Condorcet item. Hence we delete three of the 13 items so that the resulting preference matrix contains a Condorcet winner. The preference matrix is shown in Table 5.

Car dataset

This dataset contains pairwise preferences of 10 cars given by 60 users, where each car has 4 features. From the user preferences, we compute the underlying pairwise preference matrix $\mathbf{Q}$, where Q_{ij} is computed by taking the empirical average of number of times a car i is preferred over item j by the users. The preference matrix obtained in this way for Car is given in Table 6.

Sushi Dataset

This dataset contains over 100 sushis rated according to their preferences, where each sushi is represented by 7 features. Similar to Car, we construct the underlying pairwise preference matrix $\mathbf{Q}$ such that Q_{ij} is computed by taking the empirical average of number of times a sushi type i is preferred over j. We further sample 16 sushis out of these 100 such that the underlying preference matrix contains a Condorcet winner. The dataset obtained is given in Table 7.

Tennis dataset

The tennis preference matrix compiles the all-time win-loss results of tennis matches among 8 international tennis players as recorded by the Association of Tennis Professionals (ATP). For each pair of players (arms), say i and j, $\mathbf{Q}_{ij}$ is set to be the fraction of matches between i and j that were won by i. The dataset is adopted from the Tennis data used by [16] which has the item (player) 1 as its Condorcet winner. This resulted pairwise preference matrix is given below in Table 8.

0.5	0.53	0.67	0.53	0.57	0.83	0.55	0.73
0.47	0.5	0.57	0.71	0.67	0.48	0.43	0.6
0.33	0.43	0.5	0.37	0.41	0.38	0.4	0.2
0.47	0.29	0.63	0.5	0.71	0.52	0.17	0.14
0.43	0.33	0.59	0.29	0.5	0.75	0.32	0.58
0.17	0.52	0.62	0.48	0.25	0.5	0.29	0
0.45	0.57	0.6	0.83	0.68	0.71	0.5	0.52
0.27	0.4	0.8	0.86	0.42	1	0.48	0.5

Table 8: Tennis Dataset

0.5	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6
0.4	0.5	0.9	0.9	0.9	0.9	0.9	0.1	0.1	0.1
0.4	0.1	0.5	0.9	0.9	0.9	0.9	0.1	0.1	0.1
0.4	0.1	0.1	0.5	0.9	0.9	0.9	0.1	0.1	0.1
0.4	0.1	0.1	0.1	0.5	0.9	0.9	0.9	0.9	0.9
0.4	0.1	0.1	0.1	0.1	0.5	0.9	0.9	0.9	0.9
0.4	0.1	0.1	0.1	0.1	0.1	0.5	0.9	0.9	0.9
0.4	0.9	0.9	0.9	0.1	0.1	0.1	0.5	0.9	0.9
0.4	0.9	0.9	0.9	0.1	0.1	0.1	0.1	0.5	0.9
0.4	0.9	0.9	0.9	0.1	0.1	0.1	0.1	0.1	0.5

Table 5: Hurdy Dataset

0.5	0.5833	0.45	0.4333	0.6102	0.7833	0.75	0.7667	0.7119	0.7883
0.4167	0.5	0.3	0.2542	0.3051	0.5254	0.3667	0.5333	0.3729	0.5167
0.55	0.7	0.5	0.5167	0.5833	0.7627	0.7458	0.7288	0.6833	0.8136
0.5667	0.7458	0.4833	0.5	0.6333	0.7167	0.7167	0.6441	0.6724	0.7333
0.3898	0.6949	0.4167	0.3667	0.5	0.7018	0.667	0.7167	0.5763	0.7119
0.2167	0.4746	0.2373	0.2833	0.2982	0.5	0.3966	0.2881	0.35	0.4407
0.25	0.6333	0.2542	0.2833	0.333	0.6034	0.5	0.5333	0.333	0.5085
0.2333	0.4667	0.2712	0.3559	0.2833	0.7119	0.4667	0.5	0.3729	0.4576
0.2881	0.6271	0.3167	0.3276	0.4237	0.65	0.667	0.6271	0.5	0.6167
0.2117	0.4833	0.1864	0.2667	0.2881	0.5593	0.4915	0.5424	0.3833	0.5

Table 6: Car Dataset

0.5	0.705	0.534	0.72	0.533	0.429	0.591	0.398	0.683	0.626	0.528	0.554	0.66	0.573	0.534	0.575
0.295	0.5	0.392	0.643	0.299	0.32	0.42	0.198	0.489	0.416	0.304	0.39	0.493	0.458	0.393	0.376
0.466	0.608	0.5	0.729	0.5	0.451	0.522	0.295	0.62	0.503	0.383	0.52	0.634	0.572	0.485	0.53
0.28	0.357	0.271	0.5	0.262	0.268	0.255	0.191	0.432	0.373	0.222	0.305	0.425	0.346	0.312	0.259
0.467	0.701	0.5	0.738	0.5	0.462	0.559	0.224	0.68	0.556	0.381	0.487	0.676	0.575	0.513	0.56
0.571	0.68	0.549	0.732	0.538	0.5	0.635	0.336	0.728	0.605	0.508	0.545	0.703	0.666	0.515	0.603
0.409	0.58	0.478	0.745	0.441	0.365	0.5	0.317	0.615	0.481	0.359	0.482	0.637	0.556	0.441	0.419
0.602	0.802	0.705	0.809	0.776	0.664	0.683	0.5	0.794	0.683	0.683	0.652	0.743	0.738	0.663	0.697
0.317	0.511	0.38	0.568	0.32	0.272	0.385	0.206	0.5	0.453	0.299	0.36	0.437	0.371	0.343	0.306
0.374	0.584	0.497	0.627	0.444	0.395	0.519	0.317	0.547	0.5	0.476	0.478	0.558	0.517	0.466	0.476
0.472	0.696	0.617	0.778	0.619	0.492	0.641	0.317	0.701	0.524	0.5	0.553	0.739	0.675	0.566	0.627
0.446	0.61	0.48	0.695	0.513	0.455	0.518	0.348	0.64	0.522	0.447	0.5	0.621	0.608	0.524	0.553
0.34	0.507	0.366	0.575	0.324	0.297	0.363	0.257	0.563	0.442	0.261	0.379	0.5	0.347	0.335	0.273
0.427	0.542	0.428	0.654	0.425	0.334	0.444	0.262	0.629	0.483	0.325	0.392	0.653	0.5	0.447	0.419
0.466	0.607	0.515	0.688	0.487	0.485	0.559	0.337	0.657	0.534	0.434	0.476	0.665	0.553	0.5	0.504
0.425	0.624	0.47	0.741	0.44	0.397	0.581	0.303	0.694	0.524	0.372	0.447	0.727	0.581	0.496	0.5

Table 7: Sushi Dataset