Supplement material for "Reducing Dueling Bandits to Cardinal Bandits"

A. Robustness of the UCB algorithm

For completeness, we present a proof of robustness for the UCB algorithm, presented as Algorithm 1 below. Note that we did not make an effort to bound the constants in the proof. We start by presenting Chernoff's inequality providing a tail bound for estimations of variables contained in $[0,1]$.

Lemma A.1. *Let $Y_1,\dots,Y_t$ be i.i.d variables supported in $[0,1]$. Then for any $\varepsilon > 0$ it holds that*

$$\Pr\left[\frac{1}{t}\sum_{i=1}^{t} Y_i - \mathbb{E}[Y_i] > \varepsilon\right] \le e^{-2t\varepsilon^2}$$

Recall that in our setting, there are K arms, each with an expected reward. For convenience we assume the set of bandits X is the set $\{1,\dots,K\}$ and further assume for the purpose of the analysis that arm 1 has the largest expected reward. We denote by Δ_i the difference between the reward of arm 1 and that of arm i.

Proof of Lemma 2.2. For convenience, define $\beta = \alpha + 2$ where α is the robustness parameter given as input to the algorithm. For $i > 1$, define

$$u_i(t) = 2\beta\ln(t)/\Delta_i^2$$

If at time t, arm i where $i > 1$ (i.e. i is suboptimal) was chosen, one of the following must be true

1. $\rho_i(t) < u_i(t)$
2. $\hat{\mu_i} > \mu_i + \sqrt{\frac{\beta\ln(t)}{2\rho_i(t)}}$
3. $\hat{\mu_1} + \sqrt{\frac{\beta\ln(t)}{2\rho_1(t)}} < \mu_1$

Here, $\rho_i(t), \rho_1(t)$ denote the number of times arms i and 1 (the optimal arm) were pulled up to time t. Indeed, if all 3 are false we have

$$\hat{\mu_1} + \sqrt{\frac{\beta\ln(t)}{2\rho_1(t)}} \ge \mu_1 = \mu_i + \Delta_i \ge$$

$$\mu_i + 2\sqrt{\frac{\beta\ln(t)}{2\rho_i(t)}} \ge \hat{\mu_i} + \sqrt{\frac{\beta\ln(t)}{2\rho_i(t)}}$$

and the i'th arm cannot be chosen. Hence, denoting $\rho_i(T)$ the number of times arm i is queried in a total budget of T queries, we have

$$\mathbb{E}[\rho_i(T) - u_i(T)] \le \sum_{t=u_i(T)+1}^{T} \Pr[(2) \text{ or } (3)]$$

To bound the probability of event (2) occurring, we use Chernoff's inequality (Lemma A.1)

$$\Pr[(2)] \le \Pr\left[\exists\rho_i \in [t] : \hat{\mu_i} > \mu_i + \sqrt{\frac{\beta\ln(t)}{2\rho_i}}\right] \le$$

$$t \cdot t^{-\beta} = t^{1-\beta}.$$

The bound for event (3) is analogous. It follows that the probability of events (2) or (3) occurring is bounded by $2t^{1-\beta}$ and

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

$$\frac{2}{\beta-2}\left(2\beta\ln(T)\Delta_i^{-2}\right)^{2-\beta} \tag{A.1}$$

Proving the bound on the expected regret is now a matter of simple calculation

$$\mathbb{E}[R] = \sum_{i>1}\mathbb{E}[\Delta_i \cdot \rho_i(T)] \le$$

$$\frac{2\sum_i \Delta_i}{\beta-2} + \sum_{i>1} 2\beta\ln(T)/\Delta_i \le \frac{2K}{\beta-2} + 2\beta H\ln(T)$$

We proceed to prove the high probability bounds on the number of pulls of a suboptimal arm. Denote by $\rho_i^s(T)$ the number of times arm i was chosen starting from the time point $t \ge s$. Assuming $s \ge 2\beta\ln(T)\Delta_i^{-2}$, by the same arguments leading to equation A.1 we have

$$\mathbb{E}[\rho_i^s(T) - u_i(T)] \le \frac{2}{\beta-2}s^{2-\beta}$$

Assume that arm i was chosen at least s times for some

$$s \ge \frac{4(\beta+2)\ln(T)}{\Delta_i^2}$$

it follows that $\rho_i^{s-u_i(T)-1}(T) \ge u_i(T)+1$. The probability of this happening is bounded by Markov's inequality by

$$\Pr[\rho_i(T) \ge s] \le \Pr\left[\rho_i^{s-u_i(T)-1}(T) - u_i(T) \ge 1\right] \le$$

$$\mathbb{E}\left[\rho_i^{s-u_i(T)-1}(T) - u_i(T)\right] \le$$

$$\frac{2}{\beta-2}(s-u_i(T)-1)^{2-\beta} \leq \frac{2}{\beta-2}\left(\frac{s}{2}\right)^{2-\beta}$$

The last inequality holds since

$$s \geq \frac{4(\beta+2)\ln(T)}{\Delta_i^2} \geq 2 + \frac{4\beta\ln(T)}{\Delta_i^2} = 2u_i(T)+2$$

□

B. Proof of Theorem 3.1

Let $B(T)$ denote the supremum of the expected regret of the SBM S (defined in line1 of Algorithm 2) after T steps, over all possible utility distributions of the arm set X.

Fix a phase i in the algorithm. The length T_i of the phase is exactly 2^i. For all time steps t inside the phase, the left bandit x_t is drawn from some fixed distribution. Let μ' denote the common expectation $\mathbb{E}[u_t] = \mathbb{E}_{x_t}[u_t|x_t]$ of the reward of the left arm in all steps t in the phase. Now, the SBM S (defined in Line 1) is playing a standard MAB game over the set X with binary rewards. Let b_t denote the binary reward in the t'th step (within the phase). By construction,

$$\mathbb{E}[b_t|v_t, u_t] = \frac{v_t - u_t + 1}{2} \in [0,1] \,. \tag{B.1}$$

By conditional expectation, for all $y \in X$,

$$\mathbb{E}[b_t|y_t = y] = \frac{\mu(y) - \mu' + 1}{2} \in [0,1] \,. \tag{B.2}$$

Note that the arm with highest expected reward is $y = x^*$. By the definition of the bound function $B(T)$, the total expected regret (in the traditional MAB sense) of the SBM S in the phase is at most $B(T_i) = B(2^i)$. This means, that

$$\mathbb{E}\left[\sum_t \left(b_t - \frac{\mu(x^*) - \mu' + 1}{2}\right)\right] \leq B(2^i) \,,$$

where the summation runs over t in the phase. But this clearly means, using (B.2), that

$$\mathbb{E}\left[\sum_t \frac{\mu(y_t) - \mu(x^*)}{2}\right] \leq B(2^i) \,.$$

But notice that $\mathbb{E}[v_t] = \mathbb{E}_{y_t}\mathbb{E}[v_t|y_t] = \mathbb{E}[\mu(y_t)]$. Hence,

$$\mathbb{E}\left[\sum_t \frac{v_t - \mu(x^*)}{2}\right] \leq B(2^i) \,.$$

In words, this says that the expected contribution of the *right arm* to the regret (in the UDBD game) in phase i is at most $B(2^i)$. It remains to bound the expected contribution to the regret of the left bandit in phase i, which is drawn by a distribution which assigns to all $x \in X$ a probability proportional to the frequency of x played as the *right arm* in the *previous phase*.[11] By the principle of conditional expectation, and due to the linearity of the link function, the expected regret incurred by x_t (in each step in the phase) is *exactly* the average expected regret contributed by the right bandit in phase $i-1$, and hence at most $B(2^{i-1})/2^{i-1}$. This means that the total expected regret incurred by the left bandit in phase i is bounded by $2^i(B(2^{i-1})/2^{i-1}) = 2B(2^{i-1})$. Concluding, for a time horizon of T uniquely decomposable as $2+4+8+\cdots+2^k+Z$ for some integers $k \geq 1$ and $0 \leq Z \leq 2^{k+1}$-1, the total expected regret is given by the following function of T:

$$1/2 + 3B(2) + 3B(4) + \cdots + 3B(2^k) + B(Z) \,. \tag{B.3}$$

The theorem claim is now obtained by simple analysis of (B.3).

C. Proof of Theorem 4.2

To follow the proof, it is important to understand that in **MultiSBM** (Algorithm 3), exactly one SBM is advanced at each step in Line 6. This means that the internal timer of each SBM may be (and usually is) strictly behind the iteration counter of the algorithm, which is measured by the variable t. Denote by $\rho_x(t)$ the total number of times S_x was advanced after t iterations of the algorithm, for all $x \in X$.

We now assume that all coin tosses are fixed (obliviously) in advance. This allows us to discuss the regret of the SBM S_x (line 1) after T' *internal* steps even if in practice the value t for which $\rho_x(t) = T'$ might be much larger than the total number of arm pulls T , and in fact, may not even exist.

Notice that internally, S_x sees a world in which the reward is binary, and the expected reward for bandit $y \in X$ is exactly $(\mu(y) - \mu(x) + 1)/2$ at each internal step. This is because when S_x is advanced, the left bandit (in the UBDB game) is identically x. It follows that in all SBMs, the suboptimalities are the same and are $\Delta_y/2$ for arm y.

For $x \in X$ and integer $T' > 0$, let

$$R_x(T') = \frac{1}{2}\sum_{t:\rho_x(t)\leq T', x_t = x} \Delta_{y_t}$$

In words, this is the contribution of the right bandit choices to the UDBD regret at all times t for which the left bandit is chosen as x, and S_x's internal counter has not surpassed T'. The expression $R_x(T')$, by the last discussion, also

---

[11] If X is infinite, to be precise we need to say that the distribution is also supported on the set of arms played on the right side in the previous phase.

measures the expected internal regret seen by S_x after T' internal steps. Similarly, we define

$$R_{xy}(T') = \#\{t : \rho_x(t) \le T', x_t = x, y_t = y\}\Delta_y/2$$

This measures a part of $R_x(T')$ for which the right bandit is y. We start with an observation expressing the regret of the entire process as a function of the different R_{xy}'s. It will be useful to define $\rho_{xy}(T') = \#\{t : \rho_x(t) \le T', x_t = x, y_t = y\}$, so that $R_{xy}(T') = \rho_{xy}(T')\Delta_y/2$.

Observation C.1. *For any* $T \ge 1$*, the total regret* $R(T)$ *of* **MultiSBM** *after* T *steps satisfies* $\left|R(T) - 2\sum_{x\in X}\sum_{y\in X} R_{xy}(\rho_x(T))\right| \le 0.5$.

We conclude that in order to bound the expected regret $R(T)$ it suffices to bound the expressions $\mathbb{E}[R_{xy}(\rho_x(T))]$. By using the upper bound of $\rho_x(T) \le T$, we get the trivial bound for $\mathbb{E}[R(T)]$ of K times the expected regret of a single machine. The main insight is to exploit the fact that typically, $\rho_x(T)$ is order of $\ln T$ for suboptimal x. We begin with the observation that for any fixed $x, y \in X$ (x suboptimal), $s \ge 8\alpha$,

$$\begin{aligned} &\Pr[R_{xy}(T) \ge (s\ln T)/\Delta_y] \\ &\quad = \Pr[R_{xy}(T) \ge ((s/2)\ln T)/(\Delta_y/2)] \\ &\quad = \Pr[\rho_{xy}(T) \ge ((s/2)\ln T)/(\Delta_y/2)^2] \\ &\quad \le \left((s/4)\ln T)/(\Delta_y/2)^2\right)^{-\alpha} \le (s\ln T)^{-\alpha} \end{aligned} \tag{C.1}$$

This is immediate from the α-robustness of the SBM and the fact we choose $\alpha > 2$. For the same assumption on s and x, y and using the union bound,

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

We now bound the quantity $\rho_x(T)$ for any nonoptimal fixed x. Using the (trivial) fact that all $z \in X$ satisfy $\rho_z(T) \le T$, together with the fact that SBM S_x is advanced in each iteration only if x was the right bandit in the previous one, we have that for all suboptimal x,

$$\begin{aligned} &\Pr[\rho_x(T) \ge (sK\ln T)/\Delta_x^2] \\ &\le \sum_{z\in X} \Pr\left[R_{zx}(T) \ge (s\ln T)/\Delta_x\right] \le K/(s\ln T)^{\alpha}, \end{aligned} \tag{C.3}$$

where the rightmost inequality is by union bound and (C.1). Fix some $x, y \in X$ (x suboptimal). The last two inequalities give rise to a random variable Z defined as the minimal scalar for which we have

$$\forall T' \in [e, e^e, e^{e^2}, \ldots, e^{e^{\lceil \ln\ln(T)\rceil}}],$$

$$\rho_x(T) \le (ZK\ln T)/\Delta_x^2, \;\; R_{xy}(T') \le (Z\ln T')/\Delta_y$$

By (C.2)-(C.3) we have that for all $s \ge 8\alpha$, $Pr[Z \ge s] \le 2s^{-\alpha} + K(s\ln T)^{-\alpha}$. Also, conditioned on the event that $\{Z \le s\}$ we have that $R_{xy}(\rho_x(T)) \le R^s_{xy} := s \cdot e \cdot \ln((sK\ln T)/\Delta_x^2)/\Delta_y$, which is $O\left(s\Delta_y^{-1}\left(\ln\ln T + \ln K + \ln s + \ln(1/\Delta_x)\right)\right)$. Combining, $\mathbb{E}[R_{xy}(\rho_x(T))]$ is bounded above by:

$$R^{8\alpha-1}_{xy} + \sum_{i=0}^{\infty} R^{8\alpha+i}_{xy}(2(8\alpha+i)^{-\alpha} + K((8\alpha+i)\ln T)^{-\alpha}) .$$

For $\alpha = \max\{3, 2 + (\ln K)/\ln\ln T\}$, it is easy to verify that the last expression converges to $O(R^{8\alpha}_{xy})$, hence

$$\mathbb{E}[R_{xy}(\rho_x(T))] = O\left(\alpha\Delta_y^{-1}\left(\ln\ln T + \ln K + \ln(1/\Delta_x)\right)\right) .$$

Concluding, the total expected regret $\mathbb{E}[R]$ is at most $0.5 + \mathbb{E}[R_{x^*} + \sum_{x,y\in X\setminus\{x^*\}} R_{xy}]$, clearly proving the theorem.

D. Extension to more General Models

Assume the setting of Section 4. In this section we assume for simplicity that for any t and any choice of x_t, y_t, the utilities are deterministically $u_t = \mu(x_t), v_t = \mu(y_t)$. In (Yue & Joachims, 2011), the dueling bandit problem is presented where a more relaxed assumption is made on the probabilities of the outcomes of duels. Each pair of arm x, y is assigned a parameter $\Delta(x, y)$ such that the probability of x being chosen when dueling with y is $0.5 + \Delta(x, y)$. It is assumed that there exists some order $\succ$ over the arms and the Δ's hold two properties.

- *(Relaxed) Stochastic Transitivity*: For some $\gamma \ge 1$ and any pair $x^* \succ x \succ y$ we have $\gamma\Delta(x^*, y) \ge \max\{\Delta(x^*, x), \Delta(x, y)\}$.
- *(Relaxed) Stochastic Triangle Inequality*: For some $\gamma \ge 1$ and any pair $x^* \succ x \succ y$ we have $\gamma\Delta(x^*, y) \le \Delta(x^*, x) + \Delta(x, y)$.

We have analyzed **MultiSBM** (Algorithm 3) under the assumption that $\Delta(x, y) = (\mu(x) - \mu(y))/2$. It can be easliy verified that our proof holds for arbitrary Δ's under the following assumption:

- *(Relaxed) Extended Stochastic Triangle Inequality*. For some $\gamma \ge 1$, and any pair x, y (where it does not necessarily hold that $x \succ y$) it holds that $\gamma\Delta(x^*, y) \le \Delta(x^*, x) + \Delta(x, y)$.

This property is clearly held for $\Delta(x, y) = (\mu(x) - \mu(y))/2$. However, it holds for a wider family of Δ's. For example, it holds for $\Delta(x, y) = \mu(x)/(\mu(x) + \mu(y))$, assuming all μ's are in the region $[1/\gamma, 1]$. The effect of γ to the regret is given in the following theorem:

Theorem D.1. *Assume the probability for the outcome of a duel is defined according to* $\Delta(x, y)$*, where* Δ *has the Relaxed Extended Stochastic Triangle Inequality with parameter* γ*. The total expected regret of* $\mathbf{MultiSBM}$ *in the UBDB game is asymptotic to*

$$\gamma H \alpha \left(K \ln(K) + K \ln\ln(T) + \sum_{x \in X \setminus \{x^*\}} \ln(1/\Delta_x) \right) +$$

$$\ln(T) H \alpha$$

assuming the invoked MAB policy is α*-robust for* $\alpha = \max(3, \ln(K)/\ln\ln(T))$.

Notice that γ does not enter the summand of $\ln(T)$, meaning that for large values of T, the regret is unaffected by γ. We defer the proof of the theorem to the full version of the paper.

E. Proof of Observation 2.1

By definition,

$$\mathbb{E}[R_t^{\text{choice}} | (x_t, y_t)] = \mu(x^*) - \mathbb{E}[U_t^{\text{choice}} | (x_t, y_t)] \ .$$

But now note that by the definition of the link function and of U_t^{choice},

$$\mathbb{E}[U_t^{\text{choice}} | (x_t, y_t)] = \phi(u_t, v_t) u_t + \phi(v_t, u_t) v_t \geq \frac{u_t + v_t}{2}$$

where we used the assumption that for $u > v$, $\phi(u, v) > 1/2$. Now notice that the expression on the right is exactly $\mathbb{E}[U^{\text{av}} | (x_t, y_t)]$. Hence,

$$\mathbb{E}[R_t^{\text{choice}} | (x_t, y_t)] \leq \mu(x^*) - \mathbb{E}[U_t^{\text{av}} | (x_t, y_t)] = \mathbb{E}[R_t^{\text{av}}] \ .$$