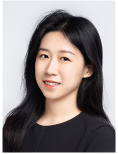
Tackling Biased Evaluators in Dueling Bandits



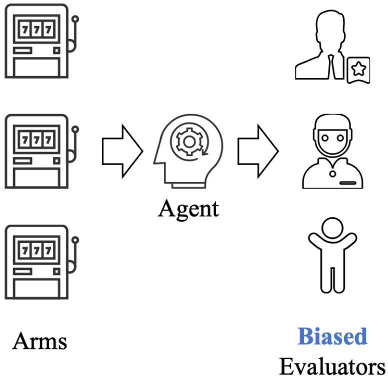
Ming Tang[*] Yuxuan Zhou[†] Chao Huang[‡]

[*]Department of Computer Science and Engineering
Southern University of Science and Technology (SUSTech), Shenzhen, China

[†]Department of Mathematics
Southern University of Science and Technology (SUSTech), Shenzhen, China

[‡]School of Computing
Montclair State University, Montclair, USA

Motivation: Dueling Bandits with Biased Evaluators



Arm Selection?

Evaluator Feedback

No Real-Valued Feedback

Pairwise Feedback by Evaluators

$\rightarrow$ Dueling Bandits

Feedback Provided by Biased Users

THIS WORK: Minimize long-term regret of arm selection

- Bias-coupled feedback;
- Unknown bias of evaluators.

System Setup

K arms:

- $p_{ij} \triangleq \Pr(o_i \succ o_j)$;
- The best arm $i^* = \arg\max_{i \in \mathcal{K}} \theta_i$, where Borda score $\theta_i \triangleq \frac{1}{K-1} \sum_{j \in \mathcal{K} \setminus \{i\}} p_{ij}$.

M biased evaluators:

- $\eta_m \triangleq \Pr(o_i \succ_m o_j \mid o_i \succ o_j)$;

In each time slot $t \in \mathcal{T} = \{1, 2, ..., T\}$:

- An arbitrary evaluator arrives;
- The agent selects two arms $\boldsymbol{x}(t) \triangleq \{x_1(t), x_2(t)\}$ for the evaluator;
- The evaluator provides pairwise comparison feedback.

System Setup

Average Regret: $\textbf{RegA}(\boldsymbol{x}(t)) = \theta_{i^*} - (\theta_{x_1(t)} + \theta_{x_2(t)})/2.$

Weak Regret: $\textbf{RegW}(\boldsymbol{x}(t)) = \theta_{i^*} - \max\{\theta_{x_1(t)}, \theta_{x_2(t)}\}.$

Goal: Minimize the long-term round-average regret:

$$\min_{[\boldsymbol{x}(t)]_{t=1}^{T}} \ \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[\textbf{Reg}(\boldsymbol{x}(t))].$$

Known Bias Case: Upper Confidence Bound (UCB)-Based Approach

For each time slot t,

- Unbiased Arm Performance Estimation:

$$\min_{\boldsymbol{p}} \sum_{i,j\in\mathcal{K},b\in\mathcal{B}_{ij}(t-1)} \left(\eta_b^{\mathrm{A}} p_{ij} + (1-\eta_b^{\mathrm{A}}) p_{ji} - \bar{p}_{ij}^{b}(t-1)\right)^2;$$

$$\hat{p}_{ij}(t-1) = \frac{\sum_{b\in\mathcal{B}_{ij}(t-1)} \left(2\eta_b^{\mathrm{A}}-1\right)\left(\bar{p}_{ij}^{b}(t-1)-(1-\eta_b^{\mathrm{A}})\right)}{\sum_{b\in\mathcal{B}_{ij}(t-1)} \left(2\eta_b^{\mathrm{A}}-1\right)^2}.$$

- Confidence Radius Calculation:

$$r_{ij}(t-1) = \frac{\sum_{b\in\mathcal{B}_{ij}(t-1)} |2\eta_b^{\mathrm{A}}-1| \sqrt{\frac{\alpha \log(t)}{(N_{ij}^{b}(t-1)+N_{ji}^{b}(t-1))}}}{\sum_{b\in\mathcal{B}_{ij}(t-1)} (2\eta_b^{\mathrm{A}}-1)^2}.$$

- Arm Selection Algorithm:

$$\mathrm{UCB}_i(t) = \frac{1}{K-1} \sum_{j\in\mathcal{K}\setminus\{i\}} \left[\hat{p}_{ij}(t-1) + r_{ij}(t-1)\right]^{-},$$

$$\max_{\boldsymbol{x}(t)} \ \mathrm{UCB}_{x_1(t)}(t) + \mathrm{UCB}_{x_2(t)}(t).$$

Known Bias Case

Theorem (Regret of Bias-Sensitive UCB Algorithm)

The bias-sensitive UCB algorithm has a round-average regret of

$$\frac{1}{T}\sum_{t=1}^{T}\mathbf{Reg}(\boldsymbol{x}(t)) \le \frac{\overline{UCB}(K(K-1)+2H)}{T} + \frac{2\overline{UCB}\sqrt{\alpha\log(T)}}{\Gamma} \times \frac{H + B^2\log(BT^{2\alpha})}{T} + \frac{2\overline{UCB}}{\Gamma}\sqrt{\frac{2\alpha K}{K-1}}\sqrt{\frac{B\log(T)}{T}}.$$

where $H = \sum_{t=K(K-1)/2+1}^{\infty} t^{-2\alpha}$, *and* $\Gamma \triangleq \sum_{b\in\mathcal{B}_{ij}(t)}(2\eta_b^{\mathrm{A}}-1)^2/|\mathcal{B}_{ij}(t)|$.

This regret is $\mathcal{O}(\sqrt{B\log(T)/T}/\Gamma)$ and hence sublinear.

Unknown Bias Case: UCB-Based Approach

For each time slot t, Arm Performance Estimation:

$$\hat{p}_{ij}(t) = \frac{\sum_{b \in \mathcal{B}_{ij}(t)} (2\hat{\eta}_m(t)-1)\left(\bar{p}_{ij}^b(t)-(1-\hat{\eta}_m(t))\right)}{\sum_{b \in \mathcal{B}_{ij}(t)} (2\hat{\eta}_m(t)-1)^2}.$$

Bias Estimation: Balancing the estimation error minimization and step size,

$$\hat{\eta}_m(t) = \arg\min_\eta \frac{1}{2}\|U_m\eta + \mathbf{b}^m\|^2 + \frac{\gamma}{2}\|\eta - \bar{\eta}_m\|^2.$$

$$\hat{\eta}_m(t) = \frac{\sum_{i,j}(2\hat{p}_{ij}(t)-1)(\bar{p}_{ij}^m(t)+\hat{p}_{ij}(t)-1)+\gamma\bar{\eta}_m}{\sum_{i,j}(2\hat{p}_{ij}(t)-1)^2+\gamma}.$$

Confidence Radius Approximation:

$$\hat{r}_{ij}(t) = \frac{\sum_{m \in \mathcal{M}_{ij}(t)} |2\hat{\eta}_m(t)-1|\sqrt{\frac{\alpha \log(t)}{(N_{ij}^m(t)+N_{ji}^m(t))}}}{\sum_{m \in \mathcal{M}_{ij}(t)} (2\hat{\eta}_m(t)-1)^2}.$$

Unknown Bias Case: UCB + Block Coordinate Descent (BCD)

Algorithm 2 Extended Bias-Sensitive UCB Algorithm

1: **Initialization:** Let $\mathcal{S}$ denote the set of $\{i, j\}$ pairs;
2: **for** each time slot $t = 1$ to $|\mathcal{S}|$ **do**
3: Select the t-th element in set $\mathcal{S}$;
4: **end for**
5: **for** each time slot $t = |\mathcal{S}| + 1$ to T **do**
6: Set $\hat{p}_{ij}(t-1) = \bar{p}_{ij}^{m}(t-1)$;
7: Compute $\hat{\eta}_m(t-1)$ using (18);
8: **Repeat** twice
9: Update $\hat{p}_{ij}(t-1)$ using (16);
10: Update $\hat{\eta}_m(t-1)$ using (18);
11: Compute $\hat{r}_{ij}(t-1)$ using (19);
12: Substitute $\hat{p}_{ij}(t-1)$ and $\hat{r}_{ij}(t-1)$ as $r_{ij}(t-1)$ into (12) and (13) to compute $\text{UCB}_i(t)$, $i \in \mathcal{K}$;
13: Select the arms $x_1(t)$ and $x_2(t)$ that optimize problem (14);
14: **end for**

Lines 6-7: Find a good starting point for updating $\hat{p}_{ij}(t-1)$ and $\hat{\eta}_m(t-1)$;

Lines 8-10: Update $\hat{p}_{ij}(t-1)$ and $\hat{\eta}_m(t-1)$ in turns for twice.

Unknown Bias Case

Definition (Parameter $\xi(t)$)

For each time slot t, let $\xi(t)$ denote the minimum non-negative value such that $\xi(t) \geq \frac{1}{2}\left(\mathrm{HF}\left(\mathbb{P}\left(|\hat{p}_{ij}(t) - p_{ij}| \leq r^{\circ}_{ij}(t)\right)\right) - \mathbb{P}\left(|\hat{p}_{ij}(t) - p_{ij}| \leq \hat{r}_{ij}(t)\right)\right)$, where $\mathrm{HF}(\cdot)$ is the tight lower bound of $\mathbb{P}\left(|\hat{p}_{ij}(t) - p_{ij}| \leq r^{\circ}_{ij}(t)\right)$.

Theorem (Regret of Extended Bias-Sensitive Algorithm)

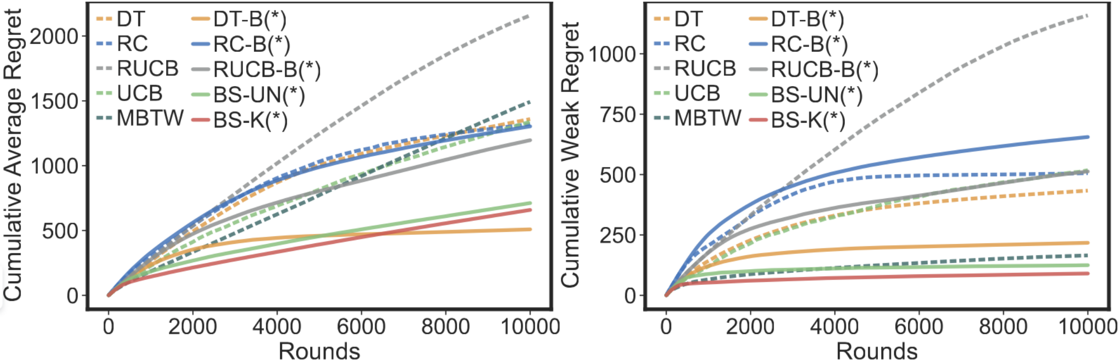
$$\frac{1}{T}\sum_{t=1}^{T} \mathbf{Reg}(\boldsymbol{x}(t)) \leq \frac{\overline{UCB}\left(K(K-1) + 2\left(H + \sum_{t=t_0}^{T} \xi(t)\right)\right)}{T} + \frac{2\overline{UCB}\sqrt{\alpha \log(T)}}{\Gamma}\left(\frac{H + B^2 \log(BT^{2\alpha})}{T} + \sqrt{\frac{2BK}{K-1} \cdot \frac{1}{\sqrt{T}}}\right).$$

Experiments

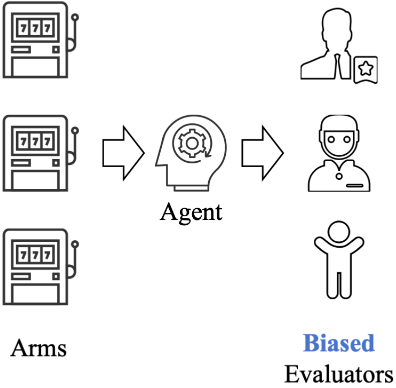
	Cumulative Average Regret (↓)						Cumulative Weak Regret (↓)					
	Arm Heter. σ^2			Bias Concentr. α_B			Arm Heter. σ^2			Bias Concentr. α_B		
	1.0	2.0	4.0	1.0	2.0	3.0	1.0	2.0	4.0	1.0	2.0	3.0
RC	1374	1338	967	2845	1338	687	596	525	502	1847	525	278
RUCB	1906	2134	1154	2832	2134	1185	1018	1144	719	1829	1144	506
DT	1396	1425	942	2621	1425	640	445	492	375	1428	492	191
MBTW	1220	1509	726	1769	1509	1448	175	162	140	569	162	92
UCB	1283	1426	732	2581	1426	706	553	548	336	1583	548	153
RC-B(*)	1378	1611	1050	2207	1611	803	649	869	727	1119	869	502
RUCB-B(*)	993	1120	709	1191	1120	1055	422	480	370	604	480	446
DT-B(*)	**430**	**411**	**344**	631	**411**	**280**	198	210	168	436	210	110
BS-UN(*)	690	689	387	825	689	637	194	161	94	340	161	92
BS-K(*)	654	713	407	**554**	713	624	**116**	**90**	**79**	**60**	**90**	**82**

- More beneficial under a higher degree of bias, i.e., smaller α_B;
- More beneficial under a lower arm heterogeneity, i.e., a smaller σ^2.

Experiments



Take-Home Messages



- Biased evaluators can lead to a significant regret increase;
- Optimization-based arm performance and bias estimation methods to mitigate regret explosion;
- Block coordinate descent for iterative estimation updates to handle nonconvex problem solving.

Welcome to contact me at tangm3@sustech.edu.cn