Dueling Bandits with Qualitative Feedback

Liyuan Xu[1,2], Junya Honda[1,2], Masashi Sugiyama[2,1]

[1]The University of Tokyo [2]RIKEN

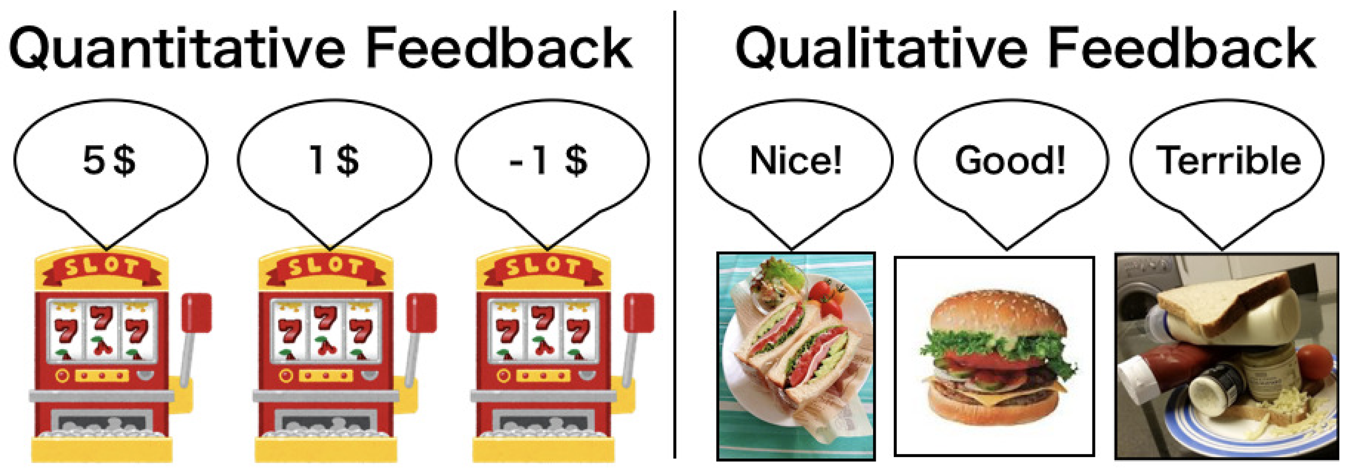
Abstract

- Dueling Bandit: The bandit problem where the best arm is defined by pairwise comparison.
- Qualitative Feedback: The feedback that can only be compared. (See problem setting.)

- Formulate a new multi-armed bandit problem that handles qualitative feedback.
- Propose algorithms reduce the same regret as the dueling bandits without explicit comparisons.
- Show the superiority of proposed algorithms theoretically and experimentally.

Problem Setting

Quantitative Feedback and Qualitative Feedback



Only qualitative feedback is available in:

- Side-effect of drugs, Quality of translated texts, Quality of results of information retrieval

Multi-armed bandits with qualitative feedback

- The set of arms $[K] = \{1, \dots, K\}$ and possible feedback in $[L]$ (the larger the better).
- An agent plays arm $a_t \in [K]$ at each round $t = 1, \dots, T$.
- Playing arm i reveals stochastic feedback $X_i \in [L] = \{1, \dots, L\}$.
- X_i follows categorical distribution $\boldsymbol{P}^{(i)}$ on $[L]$, where

$$\boldsymbol{P}^{(i)} = (P_1^{(i)}, \dots, P_L^{(i)})^\top, \; P_k^{(i)} = \mathbb{P}[X_i = k].$$

→ Since "expected reward" has no meaning, the "best arm" is unclear.

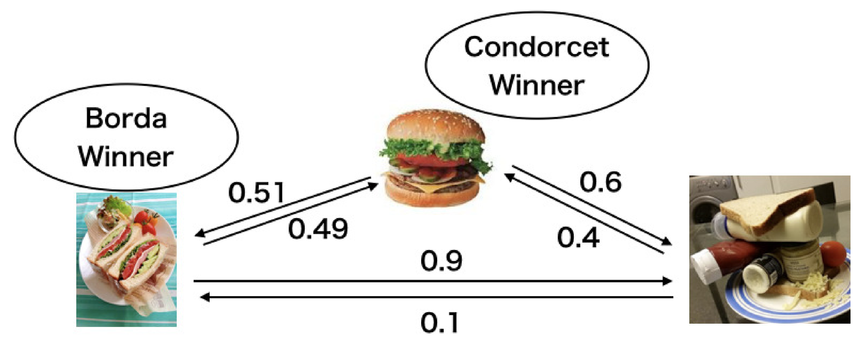
e.g. [Szorenyi+ 2015] considers τ-quantile of feedback distributions.

The Dueling Bandit

- Select two arms (i_t, j_t) at each round t, and observe the result of stochastic dueling.
- Goal: To select the winner, the arm with a high winning probability, as often as possible.

Definitions of Winners: For the winning probability $\mu_{i,j}$ of arm i over arm j,

- Condorcet winner a^*_{CW}: The arm satisfies $\forall i \neq a^*_{\mathrm{CW}}, \mu_{a^*_{\mathrm{CW}},i} \geq \frac{1}{2}$.
- Borda winner a^*_{BW}: The arm with the largest average winning probability.



In the left figure,

- Condorcet Winner is hamburger
- Borda Winner is sandwich
  average winning probability: hamburger=0.555, sandwich=0.595

The Goal of Dueling Bandits: Minimize the following regrets incurred within T rounds

- Regret of Condorcet winner:

$$R_T^{\mathrm{CW}} = \sum_{t=1}^{T} \left(\mu_{a^*_{\mathrm{CW}},i_t} - \frac{1}{2}\right) + \left(\mu_{a^*_{\mathrm{CW}},j_t} - \frac{1}{2}\right).$$

- Regret of Borda winner:

$$R_T^{\mathrm{BO}} = \sum_{t=1}^{T} (B_{a^*_{\mathrm{BO}}} - B_{i_t}) + (B_{a^*_{\mathrm{BO}}} - B_{j_t}),$$

where $B_i = \frac{1}{K-1}\sum_{j \neq i} \mu_{i,j}$ is the average winning probability.

Regret lower bound [Komiyama+ 2015, Jamieson+ 2015]

$$\liminf_{T\to\infty} \frac{R_T^{\mathrm{CW}}}{\log T} \geq \sum_{i \neq a^*_{\mathrm{CW}}} \min_{j:\mu_{i,j} \leq \frac{1}{2}} \frac{\Delta_i^{\mathrm{CW}} + \Delta_j^{\mathrm{CW}}}{d(\mu_{i,j}, 1/2)}, \quad \liminf_{T\to\infty} \frac{R_T^{\mathrm{BO}}}{\log T} \geq \frac{1}{90} \sum_{i \neq a^*_{\mathrm{BW}}} \frac{1}{(\Delta_i^{\mathrm{BW}})^2}, \qquad (1)$$

where $\Delta_i^{\mathrm{CW}} = \mu_{a^*_{\mathrm{CW}},i} - \frac{1}{2}$, $\Delta_i^{\mathrm{BW}} = B_{a^*_{\mathrm{BW}}} - B_i$, $d(x,y) = x\log\frac{x}{y} + (1-x)\log\frac{1-x}{1-y}$.

Proposed Framework: The Qualitative Dueling Bandit (QDB) Problem

At each round, play one arm a_t, and minimize the same regret as the dueling bandit

$$R_T^{\mathrm{CW}} = \sum_{t=1}^{T} \left(\mu_{a^*_{\mathrm{CW}},a_t} - \frac{1}{2}\right), \quad R_T^{\mathrm{BO}} = \sum_{t=1}^{T} (B_{a^*_{\mathrm{BO}}} - B_{a_t}),$$

where the probability $\mu_{i,j}$ that arm i wins arm j is defined as

$$\mu_{i,j} = \mathbb{P}[X_i \geq X_j] + \frac{1}{2}\mathbb{P}[X_i = X_j] \quad \Leftrightarrow \quad \mu_{i,j} = \mu(\boldsymbol{P}^{(i)}, \boldsymbol{P}^{(j)}) := \sum_{k=1}^{L} P_k^{(i)} \left(\sum_{l=1}^{k} P_l^{(j)} - \frac{1}{2} P_k^{(j)} \right).$$

Related work [Busa-Fekete+ 2013] considered this as the special instance of the dueling bandit.

- Observing feedback X_i, X_j yields accurate estimate of $\mu_{i,j}$.
- Utilizing the same algorithm as the existing algorithm to decide which arm to play.

However, if we have access to qualitative feedback, we do not have to conduct "duels"

Contribution: new algorithms without explicit comparison

Case 1: Condorcet Winner

Thompson Condorcet sampling

→ Extension of Thompson sampling

- Estimate the posterior distributions of $\boldsymbol{P}^{(i)}$ with prior of $\mathrm{Dir}(1, \dots, 1)$.
- Sample $\boldsymbol{\theta}^{(i)}$ from estimated posterior distributions.
- Play the Condorcet winner in $\{\boldsymbol{\theta}^{(i)}\}$ if it exists.
- Sample $\boldsymbol{\theta}^{(i)}$ again if the winner does not exists.

Algorithm 1: Thompson Condorcet sampling

1 Play all arms for τ_0 times each;
2 Loop $t = K\tau_0, K\tau_0 + 1, \dots$
3 Estimate posterior distributions of $\boldsymbol{P}^{(i)}$;
4 Sample $\boldsymbol{\theta}^{(i)}$ from the posterior distributions;
5 if $\exists i : \mu(\boldsymbol{\theta}^{(i)}, \boldsymbol{\theta}^{(j)}) \geq \frac{1}{2}$ *for all* $j \in [K]$ then
6 Play arm i;
7 else
8 Go to Line 4;

Theorem 1 Regret R_T^{CW} for Thompson Condorcet sampling is bounded as

$$\mathbb{E}\left[R_T^{\mathrm{CW}}\right] \leq \sum_{i \neq a^*_{\mathrm{CW}}} \frac{(1+\varepsilon)\Delta_i^{\mathrm{CW}}}{D_{\min}(\boldsymbol{P}^{(i)}, \boldsymbol{P}^{(a^*_{\mathrm{CW}})})} \log T + O\left((\log\log T)^2\right) + O\left(\frac{1}{\varepsilon^{2L}}\right),$$

where $D_{\min}(\boldsymbol{P}^{(i)}, \boldsymbol{P}^{(a^*_{\mathrm{CW}})})$ measures the gap between two distributions. It can be shown that there exists $\{\boldsymbol{P}^{(i)}\}$ which can make $d(\mu_{i,j}, 1/2)/D_{\min}(\boldsymbol{P}^{(i)}, \boldsymbol{P}^{(a^*_{\mathrm{CW}})})$ arbitrarily small.

Regret can be arbitrarily smaller than any existing dueling bandit algorithms

Case 2: Borda Winner

Thompson Borda sampling

- Similar to Thompson Condorcet sampling
- Play the Borda winner in $\boldsymbol{\theta}^{(i)}$.
- No need to re-sample since the Borda winner always exists.

Algorithm 2: Thompson Borda sampling

1 Play all arms for τ_0 times each;
2 Loop $t = K\tau_0, K\tau_0 + 1, \dots$
3 Estimate posterior distributions of $\boldsymbol{P}^{(i)}$;
4 Sample $\boldsymbol{\theta}^{(i)}$ from the posterior distributions;
5 Let $\hat{B}_i = \frac{1}{K}\sum_{j \neq i} \mu(\boldsymbol{\theta}^{(i)}, \boldsymbol{\theta}^{(j)})$;
6 Play arm $\arg\max \hat{B}_i$;

Theorem 2 There exists distributions such that regret R_T^{BW} of Thompson Borda sampling grows $\Omega(T^\alpha)$ for some $\alpha > 0$.

Thompson sampling does not always achieve $R_T^{\mathrm{BW}} = O(\log T)$

Borda-UCB

→ Extension of the UCB algorithm

- Point estimate of the average winning probability $\hat{B}_i$.
- Calculate $i_{\mathrm{UCB}} = \arg\max \hat{B}_i + \beta_i$.
- If arm i_{UCB} is the most played arm, play i_{UCB}.
- If not, play all arms other than the most played arm.

Algorithm 3: Borda-UCB

1 Pull all arms for τ_0 times each;
2 while $t \leq T$ do
3 Estimate $\boldsymbol{P}^{(i)}$ as $\hat{\boldsymbol{P}}^{(i)}$;
4 $\hat{B}_i \leftarrow \frac{1}{K-1}\sum_{k \in [K]\setminus\{i\}} \mu(\hat{\boldsymbol{P}}^{(i)}, \hat{\boldsymbol{P}}^{(k)})$;
5 $i_{\mathrm{UCB}} \leftarrow \arg\max_{i \in [K]} B_i + \beta_i$;
6 if i_{UCB} *is most played* then
7 Play i_{UCB};
8 else
9 Play all arms other than the most played arm;

Theorem 3 For appropriately chosen β_i, regret R_T^{BW} of Borda-UCB algorithm is bounded as

$$\mathbb{E}\left[R_T^{\mathrm{BW}}\right] \leq \Delta_{\mathrm{all}}^{\mathrm{BW}} \left(\frac{4\alpha}{(\Delta_{\min}^{\mathrm{BW}} - 2\varepsilon)^2} \log T + O\left(\frac{1}{\varepsilon^2}\right) \right)$$

for any $\varepsilon > 0$, where $\Delta_{\mathrm{all}}^{\mathrm{BW}} = \sum_{i \neq a^*_{\mathrm{BW}}} \Delta_i^{\mathrm{BW}}$, $\Delta_{\min}^{\mathrm{BW}} = \min_{i \neq a^*_{\mathrm{BW}}} \Delta_i^{\mathrm{BW}}$ and α is a hyper-parameter.

Borda-UCB matches the regret lower bound in the dueling bandit (1)

Experiments

MSLR-10K dataset

Information retrieval (IR) dataset which contains

- Features of a document-query pair
- User-labeled relevance (1–5)

Experimental Setting

Task: choose the best IR algorithm

At each round t:

- An agent selects a_t from 5 algorithms.
- Query q_t is sampled randomly.
- Algorithm a_t returns document d.
- The relevance of q and d is revealed as qualitative feedback.

→ The QDB problem with $K = 5$ and $L = 5$

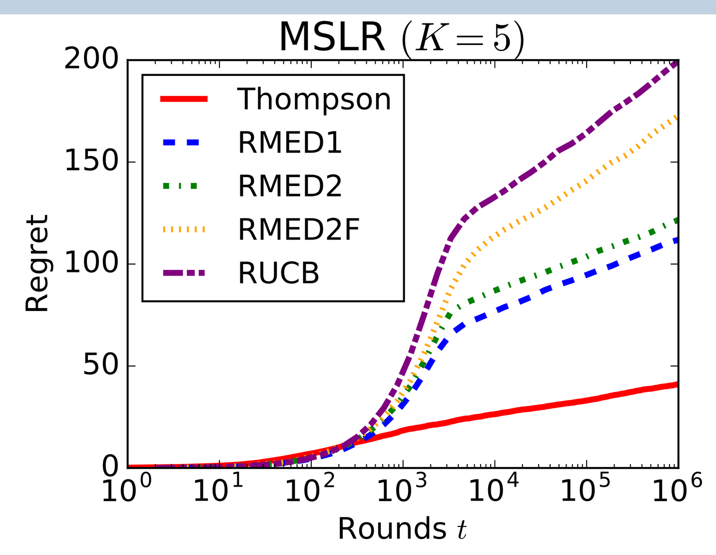


Figure 1: The experiment with the Condorcet winner

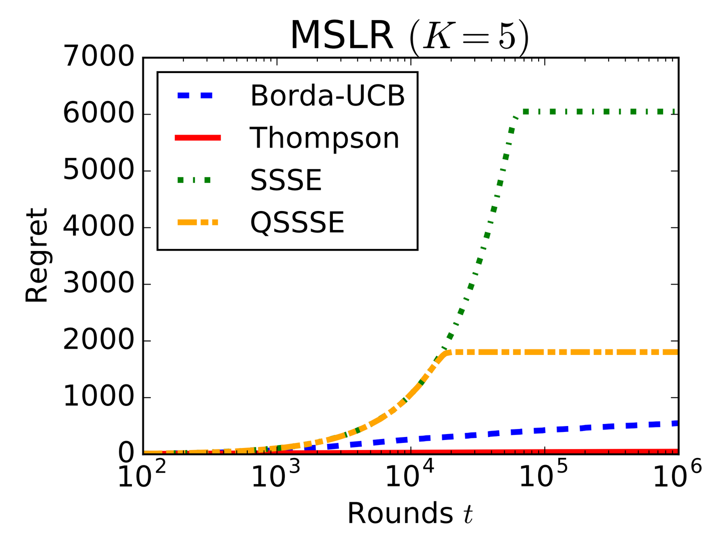


Figure 2: The experiment with the Borda winner

Vast improvement on regret compared to apply existing dueling bandit algorithms

References:

- R. Busa-Fekete, B.Szorenyi, W. Cheng, P. Weng, and E. Huellermeier. Top-k Selection based on Adaptive Sampling of Noisy Preferences. ICML2013.
- K. Jamieson, S. Katariya, A. Deshpande, and R. Nowak. Sparse dueling bandits. AISTATS2015.
- J. Komiyama, J. Honda, H. Kashima, and H. Nakagawa. Regret Lower Bound and Optimal Algorithm in Dueling Bandit Problem. COLT2015.
- B. Szorenyi, R. Busa-Fekete, P. Weng, and E. Hullermeier. Qualitative Multi-Armed Bandits: A Quantile-Based Approach. ICML2015.

THE UNIVERSITY OF TOKYO

RIKEN