Multi-Armed Bandits with Relative Feedback

Pratik Gajane

Orange labs & INRIA SequeL

7^{th} June 2017
Statistical Natural Language Processing Colloquium,
Heidelberg University

Outline

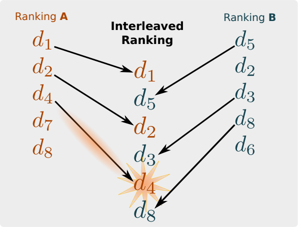
1. Dueling bandits

2. Analysis of the algorithm

3. Experiments

Motivation for the dueling bandit problem

- In many practical situations, **relative feedback** is available, and not absolute feedback.
- Eg: *"I like Tennis more than Basketball"* instead of *"I value tennis at 48/50 and Basketball at 33/50"*.



- Information retrieval systems where users provide *implicit feedback* about the provided results.
- **Interleaved filtering**, proposed by Radlinski et al. [3], interleaves the rankings to remove the bias.
- Inability of classical MAB to deal with **relative feedback** motivates new problem setting.

The dueling bandit problem

- A variation of the classical Multi-Armed Bandit (MAB) to deal with **relative feedback**.
- At each time period, the learner selects two arms.
- The learner only sees the outcome of the *duel* between the selected arms.
- The learner receives a function of the rewards of the selected arms.

Formulating the duelings bandits

- Matrix-based formulation
  - *preference matrix* contains $\mathbb{P}_{a,b}$ = unknown probability with which a wins the duel arewardst b.

$$
\begin{array}{c|cccc}
 & 1 & 2 & \cdots & K \\
\hline
\begin{array}{c} 1 \\ 2 \\ \vdots \\ K \end{array} & \multicolumn{4}{c}{\begin{bmatrix} 1/2 & \mathbb{P}_{1,2} & & \mathbb{P}_{1,K} \\ \mathbb{P}_{2,1} & 1/2 & & \mathbb{P}_{2,K} \\ & & \ddots & \\ \mathbb{P}_{K,1} & \mathbb{P}_{K,2} & & 1/2 \end{bmatrix}}
\end{array}
$$

- Utility-based formulation
  - At each time t, a utility $x_a(t)$ is associated with each arm a.
  - When arms a and b are selected,

$x_a(t) > x_b(t)$: a wins the duel

$x_a(t) < x_b(t)$: b wins the duel

$$x_a(t) = x_b(t) : \begin{cases} a \text{ wins the duel with probability } 0.5 \\ b \text{ wins the duel with probability } 0.5 \end{cases}$$

Utility-based adversarial dueling bandits

- State of the art dueling bandits algorithms are for stochastic bandits. $\rightarrow$ arm rewards are **independent and identically distributed (iid)**.
- **Adversarial** dueling bandits allow us to drop these assumptions.
- In our setting, the adversary chooses a sequence of utility vectors $\mathbf{x}(t) = (x_1(t), \ldots, x_K(t)) \in [0,1]^K$ for $t = 1, \ldots, T$.
- At each time t, the learner chooses two arms a and b,

$$\textbf{Instantaneous reward} \quad = \quad \frac{x_a(t)+x_b(t)}{2} \qquad (\textbf{hidden})$$

$$\textbf{Feedback} \quad = \quad x_a(t) - x_b(t)$$

Utility-based adversarial dueling bandits

- State of the art dueling bandits algorithms are for stochastic bandits. $\rightarrow$ arm rewards are **independent and identically distributed (iid)**.
- **Adversarial** dueling bandits allow us to drop these assumptions.
- In our setting, the adversary chooses a sequence of utility vectors $\mathbf{x}(t) = (x_1(t), \ldots, x_K(t)) \in \{0, 1\}^K$ for $t = 1, \ldots, T$.
- At each time t, the learner chooses two arms a and b,

$$\textbf{Instantaneous reward} \quad = \quad \frac{x_a(t)+x_b(t)}{2} \qquad \textbf{(hidden)}$$

$$\textbf{Feedback (binary rewards)} \quad = \quad \begin{cases} -1 & \text{if } x_a(t) < x_b(t) \\ 0 & \text{if } x_a(t) = x_b(t) \\ +1 & \text{if } x_a(t) > x_b(t) \end{cases}$$

Lower bound for any dueling bandit algorithm

Theorem

For $K \geq 2$ and $T \geq K$, there exists a distribution over assignments of rewards such that the expected cumulative regret of any utility-based dueling bandit algorithm cannot be less than $\Omega(\sqrt{KT})$.

$\mathbb{G}_{max}$ - Maximum possible reward for a single-arm strategy
$\mathbb{E}(\mathbb{G}_{alg})$ - Expected reward earned by the algorithm's strategy
$\mathbb{G}_{max} - \mathbb{E}(\mathbb{G}_{alg})$ - Expected cumulative regret

- We proved this by reduction to classical bandits as suggested in Ailon et al. [1]
- Lower bound for adversarial dueling bandits = lower bound of classical adversarial bandits $= \Omega(\sqrt{KT})$
- Data dependent lower bound for stochastic bandits $=$ $\Omega(K \log(T)/\Delta)$

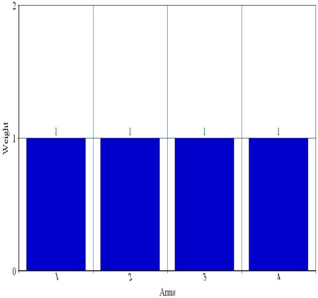
Relative Exponential Weighing Algorithm (REX3)

- Non-trivial extension of EXP3 [2] to the dueling bandits with binary rewards.
- Assigns a weight to each arm. Higher weight $\implies$ higher selection probability.
- $d = x_a - x_b = \begin{cases} -1 & \text{if } x_a < x_b \\ 0 & \text{if } x_a = x_b \\ +1 & \text{if } x_a > x_b \end{cases}$
- For anytime version, a kind of "doubling trick" (Seldin et al. [4]).

Parameters: Real $\gamma \in (0, 0.5)$
Initialization: $w_i(1) = 1$ for $i = 1, \ldots, K$.
for $t = 1, 2, \ldots$ **do**
  for $i = 1, \ldots, K$ **do**
    $p_i(t) \leftarrow (1-\gamma)\frac{w_i(t)}{\sum_{j=1}^{K} w_j(t)} + \frac{\gamma}{K}$
  end for
  Pull $a, b \sim (p_1(t), \ldots, p_K(t))$.
  Get relative feedback $d \in \{-1, 0, +1\}$
  if $a \neq b$ **then**
    $w_a(t+1) \leftarrow w_a(t) \cdot e^{\frac{\gamma}{K} \frac{d}{2p_a}}$
    $w_b(t+1) \leftarrow w_b(t) \cdot e^{-\frac{\gamma}{K} \frac{d}{2p_b}}$
  end if
  Update γ (for anytime version)

Relative Exponential Weighing Algorithm (REX3)

Weights at $t = 0$
($\gamma = 0.4$)

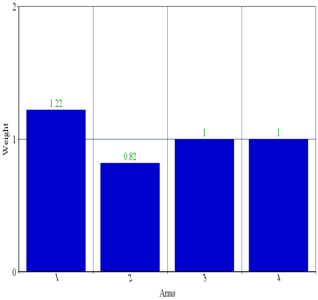


- Update weight according to (relative) feedback.

Parameters: Real $\gamma \in (0, 0.5)$
Initialization: $w_i(1) = 1$ for $i = 1, \dots, K$.
for $t = 1, 2, \dots$ **do**
    for $i = 1, \dots, K$ **do**
        $p_i(t) \leftarrow (1-\gamma)\frac{w_i(t)}{\sum_{j=1}^{K} w_j(t)} + \frac{\gamma}{K}$
    end for
    Pull $a, b \sim (p_1(t), \dots, p_K(t))$.
    Get relative feedback $d \in \{-1, 0, +1\}$
    if $a \neq b$ **then**
        $w_a(t+1) \leftarrow w_a(t) \cdot e^{\frac{\gamma}{K}\frac{d}{2p_a}}$
        $w_b(t+1) \leftarrow w_b(t) \cdot e^{-\frac{\gamma}{K}\frac{d}{2p_b}}$
    end if
    Update γ (for anytime version)

Relative Exponential Weighing Algorithm ($\mathrm{REX3}$)

$a = 1,\ b = 2,\ x_a > x_b$
Weights at $t = 1$



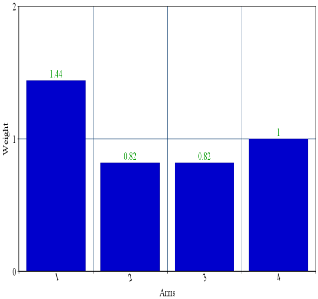
- Weight may decrease unlike $\mathrm{EXP3}$.

Parameters: Real $\gamma \in (0, 0.5)$
Initialization: $w_i(1) = 1$ for $i = 1, \ldots, K$.
for $t = 1, 2, \ldots$ **do**
  for $i = 1, \ldots, K$ **do**
    $p_i(t) \leftarrow (1-\gamma)\frac{w_i(t)}{\sum_{j=1}^{K} w_j(t)} + \frac{\gamma}{K}$
  end for
  Pull $a, b \sim (p_1(t), \ldots, p_K(t))$.
  Get relative feedback $d \in \{-1, 0, +1\}$
  if $a \neq b$ **then**
    $w_a(t+1) \leftarrow w_a(t) \cdot e^{\frac{\gamma}{K}\frac{d}{2p_a}}$
    $w_b(t+1) \leftarrow w_b(t) \cdot e^{-\frac{\gamma}{K}\frac{d}{2p_b}}$
  end if
  Update γ *(for anytime version)*

Relative Exponential Weighing Algorithm ($\mathrm{REX3}$)

$a = 1,\ b = 3,\ x_a > x_b$

Weights at $t = 2$



- Weights spike at arms who win the duel regularly.

Parameters: Real $\gamma \in (0, 0.5)$

Initialization: $w_i(1) = 1$ for $i = 1, \ldots, K$.

for $t = 1, 2, \ldots$ **do**

    for $i = 1, \ldots, K$ **do**

        $p_i(t) \leftarrow (1-\gamma)\frac{w_i(t)}{\sum_{j=1}^{K} w_j(t)} + \frac{\gamma}{K}$

    end for

    Pull $a, b \sim (p_1(t), \ldots, p_K(t))$.

    Get relative feedback $d \in \{-1, 0, +1\}$

    if $a \neq b$ **then**

        $w_a(t+1) \leftarrow w_a(t) \cdot e^{\frac{\gamma}{K}\frac{d}{2p_a}}$

        $w_b(t+1) \leftarrow w_b(t) \cdot e^{-\frac{\gamma}{K}\frac{d}{2p_b}}$

    end if

    Update γ (for anytime version)

Upper bound for REX3

Theorem

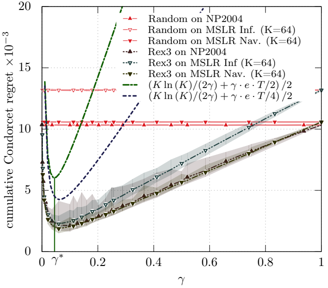
$\mathbb{G}_{max} - \mathbb{E}(\mathbb{G}_{alg}) \leq \frac{K}{\gamma}\ln(K) + \gamma\tau$

where

$\tau = e \cdot \mathbb{E}\mathbb{G}_{alg} - (4-e) \cdot \mathbb{E}\mathbb{G}_{unif}$

Corollary

When $\gamma = \min\left\{\frac{1}{2}, \sqrt{\frac{K\ln(K)}{\tau}}\right\}$, *the expected cumulative regret of* REX3 *is bounded by* $\mathcal{O}\left(\sqrt{K\ln(K)T}\right)$.



- Upper bound of REX3 = Upper bound of EXP3.
- Optimality: REX3 $\sim_{\ln} \Omega\left(\sqrt{KT}\right)$

Analysis of $\mathrm{REX3}$

- Main challenge of dueling bandits: no direct way to estimate absolute reward values like $\mathrm{EXP3}$.
- In $\mathrm{EXP3}$, since we can observe absolute feedback (x_a), the estimator $\hat{x}_i(t)$ is defined as follows:

$$\hat{x}_i(t) = [\![i = a]\!] \frac{x_a(t)}{p_a(t)}$$

- The division by p_a ensures that more *"surprising"* (i.e. lower p_a) the observed reward x_a, higher is the estimator.
- Ensures that their expectations are equal to the actual rewards for each action i.e.

$$\mathbb{E}[\hat{x}_i(t)] = x_i(t)$$

Analysis of REX3

- Feedback in dueling bandits is relative $(x_a - x_b)$ instead of absolute (x_a), so the use of EXP3 estimator is not possible.
- To overcome this challenge, we introduced a new estimator $\hat{c}_i(t)$.
- We define $\hat{c}_i(t)$ in the following way:

$$\hat{c}_i(t) = [\![i = a]\!] \frac{(x_a - x_b)}{2p_a} + [\![i = b]\!] \frac{(x_b - x_a)}{2p_b}$$

- It gives us a way to provide weight update rule in a concise form:

Weight update rule earlier

10: $w_a(t+1) \leftarrow w_a(t) \cdot e^{\frac{\gamma}{K} \frac{d}{2p_a}}$

11: $w_b(t+1) \leftarrow w_b(t) \cdot e^{-\frac{\gamma}{K} \frac{d}{2p_b}}$

Weight update rule using $\hat{c}_i(t)$ $\quad \forall i \; w_i(t+1) = w_i(t) \cdot e^{\frac{\gamma}{K} \hat{c}_i(t)}$

Key element of the analysis

Lemma for expectation of $\hat{c}_i(t)$

$$\mathbb{E}\left[\hat{c}_i(t) | (a_1, b_1), .., (a_{t-1}, b_{t-1})\right] = x_i(t) - \mathbb{E}_{a \sim p(t)} x_a(t)$$

- The expectation of this estimator is the expected instantaneous regret of the algorithm against arm i.
- *i.e.* the difference between the gain of arm i and the expected gain according to algorithm's current state of knowledge $p(t)$.
- This is intuitively what we want from an estimator in a dueling bandit problem.

Sketch of proof

The general structure of the proof is similar to the proof of EXP3 [2] except the difference in expectation of the $\hat{c}_i(t)$ estimator.
Let $W_t = w_1(t) + w_2(t) + \cdots + w_K(t)$.

$$\frac{W_{t+1}}{W_t} = \sum_{i=1}^{K} \frac{p_i(t) - \gamma/K}{1-\gamma} e^{(\gamma/K)\hat{c}_i(t)} \tag{1}$$

As in EXP3, we simplify, take the logarithm and sum over t. We get for any j:

$$\sum_{t=1}^{T} \frac{\gamma}{K} \hat{c}_j(t) - \ln(K) \leq \frac{\gamma^2/K}{1-\gamma} M_1 + \frac{(e-2)\gamma^2/K}{1-\gamma} M_2$$

Sketch of proof (continued)

By taking the expectation over the algorithm's randomization, we obtain for any j:

$$\sum_{t=1}^{T} \frac{\gamma}{K} \mathbb{E}_{\sim p} \hat{c}_j(t) - \ln(K) \leq$$

$$\frac{\gamma^2/K}{1-\gamma} \sum_{i=t}^{T} \mathbb{E}_{\sim p} M_1 + \frac{(e-2)\gamma^2/K}{1-\gamma} \sum_{i=t}^{T} \mathbb{E}_{\sim p} M_2 \qquad (2)$$

Sketch of proof (continued)

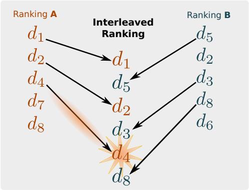
From Lemma 13, expectation of M_1, expectation of M_2, and by definition of $\mathbb{G}_{max}$, $\mathbb{E}\mathbb{G}_{alg}$, and $\mathbb{E}\mathbb{G}_{unif}$, the inequality (2) rewrites into:

$$\mathbb{G}_{max} - \mathbb{E}\mathbb{G}_{alg} - \frac{K \ln K}{\gamma} \leq \frac{\gamma}{1-\gamma} \left(\mathbb{E}\mathbb{G}_{alg} - \mathbb{E}\mathbb{G}_{unif} \right)$$
$$+ \frac{(e-2)\gamma}{2(1-\gamma)} \left(\mathbb{E}\mathbb{G}_{alg} + \mathbb{E}\mathbb{G}_{unif} \right)$$

Assuming $\gamma \leq \frac{1}{2}$ we finally obtain:

$$\mathbb{G}_{max} - \mathbb{E}\mathbb{G}_{alg} \leq \frac{K \ln K}{\gamma} + \gamma \left(e\mathbb{E}\mathbb{G}_{alg} - (4-e)\, \mathbb{E}\mathbb{G}_{unif} \right)$$

Experiments



- We used **interleaved filtering** on **real datasets** from information retrieval systems.
- We considered the following state of the art algorithms: BTM [6] (explore-then-exploit setting), SAVAGE [5], RUCB [7], and SPARRING coupled with EXP3 [1] and Random as baseline.
- The experiments showed that REX3 and especially its anytime version are competitive solutions for the dueling bandit problem.

Experiments

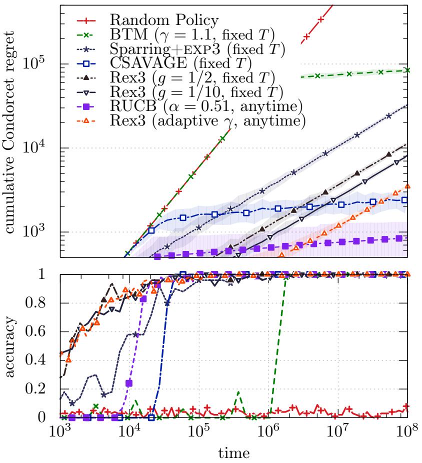


Figure 1: Average regret and accuracy plots on ARXIV dataset (6 rankers). Time and regret scales are logarithmic.

Experiments

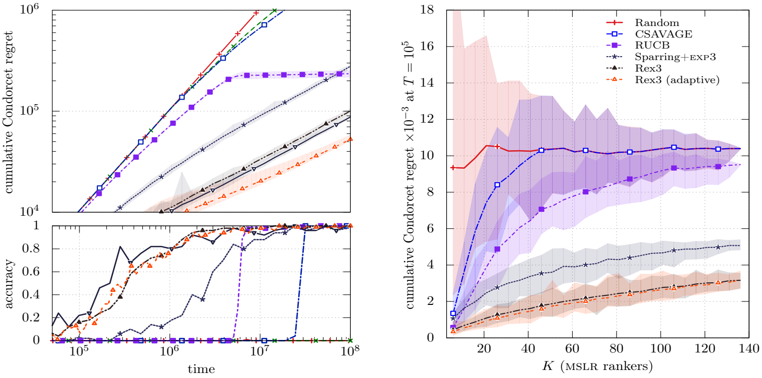


Figure 2: On the left: average regret and accuracy plots on MSLR30K with navigational queries ($K = 136$ rankers). On the right: same dataset, fixed $T = 10^5$ and $K = 4$ - 136. Colored areas show minimal and maximal values.

Simulations on non-stationary rewards

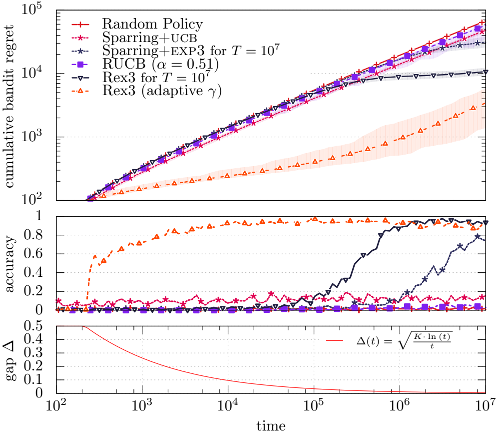


Figure 3: $K = 10$, gains from Bernoulli distributions. Best arm's gain is $1/2 + \Delta(t)$ with $\Delta(t) = \sqrt{K \cdot \log(t)/t}$. Others are stationary with a mean of $1/2$.

References I

N. Ailon, Z. Shay Karnin, and T. Joachims.
Reducing dueling bandits to cardinal bandits.
In *ICML 2014*, volume 32 of *JMLR Proceedings*, pages 856–864, 2014.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire.
The nonstochastic multiarmed bandit problem.
SIAM Journal on Computing, 32(1):48–77, 2002.

F. Radlinski and T. Joachims.
Active exploration for learning rankings from clickthrough data.
In *KDD 2007*, pages 570–579. ACM, 2007.

References II

Yevgeny Seldin, Csaba Szepesvári, Peter Auer, and Yasin Abbasi-Yadkori.
Evaluation and analysis of the performance of the exp3 Algorithm in stochastic environments.
In *EWRL*, volume 24 of *JMLR Proceedings*, pages 103–116, 2012.

T. Urvoy, F. Clerot, R. Féraud, and S. Naamane.
Generic exploration and K-armed voting bandits.
In *ICML 2013*, volume 28 of *JMLR Proceedings*, pages 91–99, 2013.

Y. Yue and T. Joachims.
Beat the mean bandit.
In *ICML 2011*, pages 241–248. Omnipress, 2011.

References III

Masrour Zoghi, Shimon Whiteson, Remi Munos, and Maarten de Rijke.
Relative upper confidence bound for the k-armed dueling bandit problem.
In *ICML 2014*, volume 32 of *JMLR Proceedings*, pages 10–18, 2014.