*Probability in the Engineering and Informational Sciences*, **36**, 2022, 264–275.
doi:10.1017/S0269964820000601

# DUELING BANDIT PROBLEMS

EROL PEKÖZ
*Boston University, Boston, MA, USA*
*E-mail: pekoz@bu.edu*

SHELDON M. ROSS AND ZHENGYU ZHANG
*University of Southern California, Los Angeles, CA, USA*
*E-mails: smross@usc.edu; zhan892@usc.edu*

There is a set of $n$ bandits and at every stage, two of the bandits are chosen to play a game, with the result of a game being learned. In the "weak regret problem," we suppose there is a "best" bandit that wins each game it plays with probability at least $p > 1/2$, with the value of $p$ being unknown. The objective is to choose bandits to maximize the number of times that one of the competitors is the best bandit. In the "strong regret problem", we suppose that bandit $i$ has unknown value $v_i$, $i = 1, \ldots, n$, and that $i$ beats $j$ with probability $v_i/(v_i + v_j)$. One version of strong regret is interested in maximizing the number of times that the contest is between the players with the two largest values. Another version supposes that at any stage, rather than choosing two arms to play a game, the decision maker can declare that a particular arm is the best, with the objective of maximizing the number of stages in which the arm with the largest value is declared to be the best. In the weak regret problem, we propose a policy and obtain an analytic bound on the expected number of stages over an infinite time frame that the best arm is not one of the competitors when this policy is employed. In the strong regret problem, we propose a Thompson sampling type algorithm and empirically compare its performance with others in the literature.

**Keywords:** applied probability, simulation, stochastic modeling

## 1. INTRODUCTION

In the classical stochastic multi-armed bandit problem (MAB), in each time period, the learner selects an arm from a given set of $n$ arms and then observes a random reward for that selected arm. The goal of the learner is to minimize the cumulative regret, defined as the expected difference between the sum of rewards that could have been received by always playing the best arm and the sum of rewards actually received. In this paper, we study the dueling bandit problem, a variant of classical MAB, where the action is to compare a pair of arms rather than pulling one single arm. More specifically, at each time step, the learner chooses 2 arms and then observes which of the two arms is preferred (or, equivalently, which arm is the winner of a duel between these arms). The problem arises naturally in domains

 0269-9648/19 $25.00 

where feedback is represented in the form of pairwise comparison, such as recommendation systems, ad replacements, and information retrieval.

There are two types of optimality criteria that are generally considered in dueling bandits: weak regret and strong regret. Weak regret is concerned with minimizing the expected number of times that the best arm is not one of the two arms selected. In this paper, for the weak regret problem, we assume that there is an unknown arm $i^*$ and an unknown probability $p > 1/2$, such that, independent of anything that came earlier, arm $i^*$ is the winner of each of its duels with probability at least $p$. (Thus, we assume nothing about the probabilities concerning games not involving $i^*$, nor even whether such games are conditionally independent given their players.) In Section 2, we propose an algorithm, called Beat the Winner (BTW) and show that the expected weak regret over an infinite time horizon is bounded by an $O(n^2)$, where $n$ is the number of arms. We also propose a modification of this rule, called Modified Beat the Winner (MBTW) and empirically show that it has a smaller regret than both BTW and another algorithm, called WS-W, recently considered in the literature (see [2]). We also show in this section that the analytic bound for the expected infinite horizon weak regret of WS-W given in Chen and Frazier [2] can be improved.

We consider two types of strong regret. In strong regret 1, not previously considered in the literature, we suppose that the objective is to maximize the expected number of times that the two best arms (to be defined) are chosen. In strong regret 2, it is supposed that the two arms in the duel may be the same arm and the objective is to maximize the expected time that the best arm is chosen as both of the dueling arms. (A more natural description of strong regret 2 is that at any time, rather than choosing two arms to duel, one can make a declaration that a specified arm is the best.) We consider the strong regret problem under the assumption that arm $i$ has an unknown associated value $v_i,\ i = 1, \ldots, n$ and the probability of arm $i$ beating arm $j$ is $v_i/(v_i + v_j)$. In Section 3, we present a Thompson sampling algorithm, which utilizes an MCMC simulation approach to sample associated values of arms from the posterior distribution, with these sampled values then used to decide which arms to next pair. Empirical evaluation is made to compare our procedure with others in the literature.

Dueling bandit problem was originally raised by Yue *et al.* [8] and has been primarily studied under the strong regret 2 criterion. Various definitions of the best arm (winner) have been considered. Early algorithms, such as IF ([8]) and BTM ([]), assumed that $i$ is the winner over $j$ with unknown probability $p_{ij},\ i \neq j$, and supposed that the arms are totally ordered in that for some permutation $i_1, \ldots, i_n$ of $1, \ldots, n$, $p_{i_j, i_k} > 0.5$ when $j > k$. Both algorithms adopted an exploration then exploitation strategy to control the regret. As an extension, the Condorcet winner model only assumes that there exists an arm who beats any other arm with probability greater than 0.5. Zoghi *et al.* [10] proposed RUCB algorithm by adopting the UCB framework and provided the theoretical guarantee that the cumulative regret is upper bounded by $O(n^2 \log m)$ where $m$ is the time horizon. The later proposed merge RUCB in Zoghi *et al.* [12] further tightened the regret upper bound to $O(n \log m)$. Other algorithms including RMED ([3]) and WS-S ([2]) also achieved $O(n \log m)$ regret upper bound. Beyond the Condorcet winner setting, some other notions of winner, such as Copeland winner and Borda score, have also attracted much attention ([1,4,6,12]). In our work, we consider the model under the strong regret where each arm has an associated value and those values explicitly specify the winning probability.

Dueling bandits under the weak regret criterion have also previously been considered ([2,8]). To the best of our knowledge, the recent work of Chen and Frazier [2] seems to be the only paper that studied the weak regret and designed a specific algorithm (called WS-W) for it. A brief description of WS-W is as follows. Let the score of each arm be the number of wins minus the number of losses in all games that arm has played. Round $k + 1,\ k \geq 0$, will

https://doi.org/10.1017/S0269964820000601 Published online by Cambridge University Press

begin with one arm having score $(n-1)k$ and all others having score $-k$. The player with score $(n-1)k$ will play with a randomly chosen one of the other arms a series of games that ends when one of their scores is $-k-1$. At that point, the player with score $-k-1$ stops playing in round $k+1$ and the other plays with a randomly chosen one of the remaining arms until one of their scores hits $-k-1$, and so on. It was shown in Chen and Frazier [2] that WS-W has $O(n^2)$ bound under the Condorcet winner setting and $O(n \log n)$ if arms are totally ordered.

There are various applications of dueling bandits. For instance, content recommendation systems, such as the in-app restaurant recommendations of Grubhub and UberEATS, have a goal of learning user preferences and presenting a short list of personalized recommendations. The algorithm learns user preferences over a large set of items from the choices users make when presented with pairs of recommendations. After each pair of items is presented, we can observe the one that is preferred by the user, and the algorithm must choose the next pair of items to present. This online recommendation problem can be cast as our dueling bandit problem where the item corresponds to arms and the goal is to recommend customers' favorite item.

## 2. WEAK REGRET: BEAT THE WINNER

In this section, we propose two algorithms, named *Beat the winner* (BTW) and *Modified Beat the winner* (MBTW) for dueling bandits in the weak regret criterion. We first introduce our model assumptions. Suppose that there is a set of $n$ arms. At each time step, the learner chooses two arms to play a game and observes the winner. We let $p_{i,j}$ denote the probability that arm $i$ beats arm $j$ when $i$ plays with $j$. We assume that there exists a Condorcet winner, that is, there is an unknown arm $i^*$ and an unknown probability $p > 1/2$, such that arm $i^*$ is the winner of each its duels with probability at least $p$. We call $i^*$ the best arm. The binary weak regret $r(t)$ at time period $t$ is $r(t) = 0$ if the best arm is one of the chosen arms and $r(t) = 1$ otherwise. Our objective is to minimize the cumulative weak regret $\sum_{t=1}^{\infty} r(t)$ over the infinite time horizon.

### 2.1. Beat the Winner

We now present our BTW algorithm. The BTW algorithm proceeds in rounds, with round $k, k \geq 1$, consisting of two arms playing a sequence of games until one of them has won $k$ times.

*Beat the Winner Rule*

- Arms are initially put in the queue in a random order.
- For round $k = 1, 2, \ldots$.
  - Top two arms in queue play a sequence of games. The winner of the round is the first to win $k$ games.
  - The loser goes to the end of the queue, and the winner stays at the top of the queue.

Now we show the expected cumulative regret over an infinite time horizon of BTW is upper bounded by $O(n^2)$.

LEMMA 2.1: *Let $L_k$ be the event that the best arm $i^*$ is the loser of round $k$. With $a = 2p - 1$*

$$\mathbb{P}(L_k) \leq \exp\{-ka^2\}$$

https://doi.org/10.1017/S0269964820000601 Published online by Cambridge University Press

PROOF: Because $i^*$ must play in round $k$ to be the loser of that round,

$$\mathbb{P}(L_k) \leq \mathbb{P}(L_k \,|\, i^* \text{ plays in round } k\}$$

Now, it follows by a coupling argument that $\mathbb{P}(L_k|i^* \text{ plays in round } k\}$ is upper bounded by the probability that a total of $k$ heads occurs before a total of $k$ tails in a sequence of independent trials that each results in a head with probability $q = 1 - p$. Hence, with $B$ being a binomial $(2k-1, q)$ random variable,

$$\begin{aligned}\mathbb{P}(L_k) &\leq \mathbb{P}(B \geq k) \\ &= \mathbb{P}(B - (2k-1)q \geq ka + q) \\ &\leq \exp\left\{-\frac{2(ka+q)^2}{2k-1}\right\} \\ &\leq \exp\{-ka^2\}\end{aligned}$$

where the second inequality follows from Chernoff's bound. ■

THEOREM 2.2: *With $X$ being the total number of games that do not involve $i^*$,*

$$E[X] \leq (n-2)^2 + e^{-a^2}(2K^3 + (2n-5)K^2)$$

*where $K = 1/(1 - e^{-a^2})$.*

PROOF: Let $R$ be last round lost by $i^*$. Lemma 2.1 gives

$$\mathbb{P}(R \geq r) = \mathbb{P}\left(\bigcup\nolimits_{k \geq r} L_k\right) \leq K \exp(-ra^2)$$

Hence,

$$E[R] = \sum_{r \geq 1} \mathbb{P}(R \geq r) \leq K^2\, e^{-a^2}$$

and

$$\begin{aligned}E[R^2] &= 2\sum_{r \geq 1} r\, \mathbb{P}(R \geq r) - E[R] \\ &\leq 2K^3\, e^{-a^2} - E[R]\end{aligned}$$

Because there are at most $2k-1$ games in round $k$, and $i^*$ plays in all rounds after round $R + n - 2$

$$X \leq \sum_{i=1}^{R+n-2} (2i-1) = (R+n-2)^2$$

yielding that

$$E[X] \leq (n-2)^2 + e^{-a^2}(2K^3 + (2n-5)K^2) \qquad \textbf{(2.1)}$$

■

Note that the regret bound of the BTW matches that of the WS-W proposed in Chen and Frazier [2] without assuming $p_{ij} \neq 0.5$ for any pair $i, j$ (We will show later that WS-W actually does not need that assumption.) However, in practice, BTM is not very competitive

https://doi.org/10.1017/S0269964820000601 Published online by Cambridge University Press

for small $n$ since the BTM takes relatively long time to identify and extensively play the best arm. This is also indicated by the regret bound derived in Eq. (2.1), where $n^2$ is dominated by the constant $K^3$ when $n$ is small. On the other hand, when $n$ is large, the performance of BTW roughly matches WS-W and enjoys the advantage of having a smaller variance. Numerical instances will be shown in the next section along with our proposed MBTW algorithm.

### 2.2. Modified Beat the Winner

Note that one of the main drawbacks of the BTW is that it does not utilize any past records of arms. Hence, two apparently bad arms could play a large number of games, where we gain no meaningful information. To overcome this drawback, we consider allowing the learner to keep track of some records of arms. Specifically, for each arm, we want to record the difference between the number of rounds an arm wins and the number of rounds it loses. Based on such information, we propose the MBTW algorithm and empirically show that it significantly outperforms both BTM and WS-W.

Similar to BTW, MBTW also plays games in a round fashion, where each round consists of a series of games. However, the number of games is no longer determined by the number of past rounds, but by the records of the arms in the duel. We now show how to define the records of the arms, and how to choose the players of the next round.

*Modified Beat the Winner Rule*

- The initial value of $r_i$, the score of arm $i$, is $r_i = 1, i = 1, \ldots n$
- Choose an arm uniformly at random as the *host*, and let $h$ denote the index of the host. The current host is always one of the players of the next round.
- For each round
  - Let $i, i \neq h$, be the opponent of arm $h$ with probability $r_i / \sum_{i \neq h} r_i$. The arm $i$ so chosen and $h$ play a sequence of games until one of them has won $r_h$ games. Let $w$ and $l$ denote the indices of the winner and loser.
  - Reset $r_w = r_w + 1$
  - Reset $r_l = \max(r_l - 1,\ 1)$
  - Set $h = w$. (The winner of the current round becomes the host.)

Two simulated numerical instances with 100 arms and 1,000 arms are shown in Figures 1 and 2, including the plot of cumulative regret and standard deviation at fixed time horizon $T$. The WS-W method is used as the benchmark to evaluate our proposed algorithms. In each case, the simulated results are based on 2,000 simulation runs, where each run begins with generating the probabilities $p_{i,j}$ from uniform $(0.2, 0.8)$ random variables for $i < j,\ i, j \neq i^*$, and generating $p_{i^*,j}$ from uniform $(0.5, 0.8)$ for the 100 arms case and from uniform $(0.55, 0.8)$ for the 1000 arms case. Note that we increase the lower bound of the winning probability of the Condorcet winner in the second instance so as to speed up the convergence.

### 2.3. A Revisit to WS-W

For the rest of this section, we provide a supplementary proof of WS-W showing that the upper bound of expected cumulative regret under the Condorcet winner setting can be further improved over what was shown in Chen and Frazier [2]. Compared to the original proof, our proof is still valid when there exists a pair $i, j$ with $p_{i,j} = 0.5$ and thus regret

https://doi.org/10.1017/S0269964820000601 Published online by Cambridge University Press

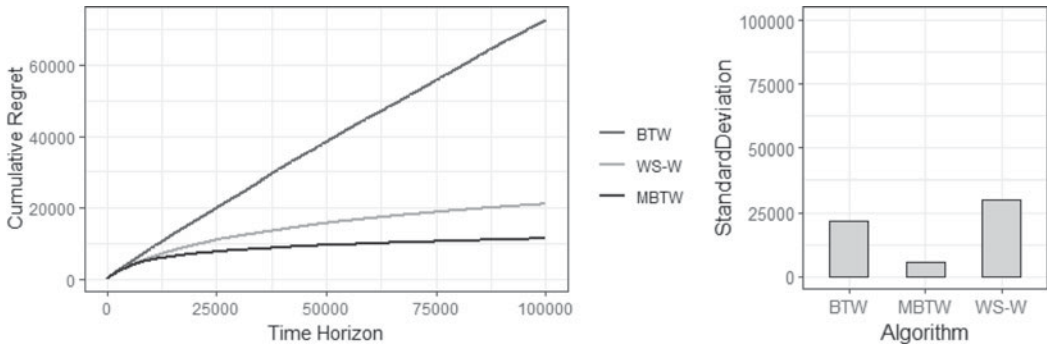


FIGURE **1.** Experiments with 100 arms. (a) Cumulative regret over 100 replications. (b) Standard deviation of cumulative regret for $T = 10^5$.

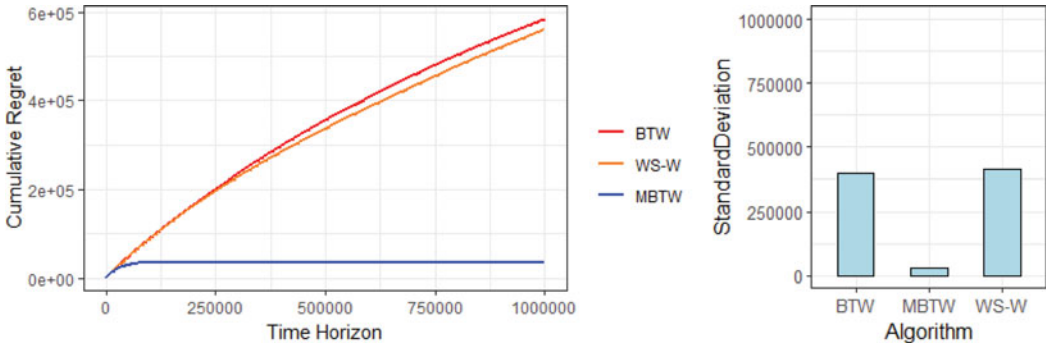


FIGURE **2.** Experiments with 1000 arms. (a) Cumulative regret over 100 replications. (b) Standard deviation of cumulative regret for $T = 10^6$.

bound only depends on the smallest winning probability of the Condorcet player when matched with the other players.

Now consider the WS-W algorithm. Let $W_k$ and $L_k$ be the event that the best arm wins round $k$ and loses round $k$, respectively. To slightly simplify the following analysis, we assume that there are $n + 1$ arms in total.

LEMMA 2.3: *Condition on whether the best arm is the winner of round $k - 1$, the probability that the best arm wins round $k$ is*

$$P(W_k \mid W_{k-1}) \geq \frac{1 - (q/p)^{nk-n+k}}{1 - (q/p)^{n(k+1)}}$$

$$P(W_k \mid L_{k-1}) \geq \frac{1 - (q/p)}{1 - (q/p)^{n(k+1)}}$$

LEMMA 2.4: *The probability that the best arm loses round $k$ is bounded by*

$$P(L_k) < 2\left(\frac{q}{p}\right)^k$$

https://doi.org/10.1017/S0269964820000601 Published online by Cambridge University Press

LEMMA 2.5: *Consider gambler's ruin problem which stops when the gambler is either up* $m-1$ *or down* 1. *Let* $E_p[X]$ *be the mean number of games when the gambler wins each bet with probability* $p$. *Then for* $\forall p \in (0,1)$

$$E_p[X] < 2m$$

The proofs of Lemmas 2.3, 2.4 and 2.5 can be found in the Appendix.

THEOREM 2.6: *The expected cumulative regret of WS-W is bounded by* $(2p^2/(2p-1)^2)(n^2+n)$

PROOF: Let $X$ be the total cumulative regret over infinite time horizon. If we let $r = q/p$, then

$$\begin{aligned} E[X] &= \sum_{k\geq 1} E[\text{regret at round k}] \\ &\leq \sum_{k\geq 1} 2r^{k-1}k(n^2+n) \\ &= \frac{2p^2}{(2p-1)^2}(n^2+n) \end{aligned}$$

where the inequality follows by Lemmas 2.3 and 2.4. ■

## 3. STRONG REGRET: THOMPSON SAMPLING APPROACH

In this section, we restrict ourselves to the scenario where each arm has an unknown associated value $v_i,\ i = 1,\ldots,n$. The probability that arm $i$ is preferred over $j$ is $v_i/(v_i+v_j)$. The objective is to minimize two versions of cumulative strong regret. Specifically, under the notion of strong regret 1, two different arms are picked at each time slot and one can avoid the regret only if the two best arms (i.e. two arms with largest $v_i$) are selected. On the other hand, under the strong regret 2 model, one is allowed to pick the same arm in the duel. The strong regret 2 objective then is to minimize the number of times that the best arm is not chosen as both the dueling arms. Note that we use binary regret under both settings, meaning that regret is 0 if the corresponding optimal criteria are satisfied and 1 otherwise.

We propose a new algorithm by adopting the Thompson sampling approach, originally introduced by Thompson [5] for the classical MAB problem. The existing works that employ the Thompson sampling idea on dueling bandits start by assuming that the quantities $p_{i,j}, i<j$ are the values of $\binom{n}{2}$ independent uniform $(0,1)$ random variables and define the best arm as the $i$ that maximizes $\sum_{j\neq i} p_{i,j}$. In this manner, the posterior random variables are independent beta random variables, and so are easy to simulate from (see RCB [11] and DT [7]). In our model, however, the preference probability is completely determined by the associated values and thus the joint posterior distribution of the values is no longer independent. In the following work, we develop an MCMC sampling approach that allows us to sample values of arms from the posterior distribution. We then empirically compare our algorithm to five benchmarks.

https://doi.org/10.1017/S0269964820000601 Published online by Cambridge University Press

### 3.1. The Sampling Approach

The approach of sampling values of arms at each time stage is as follows. We imagine that the unknown values $v_1, \ldots, v_n$, are the values of independent mean 1 exponential random variables $V_1, \ldots, V_n$. Given this, it follows that if $w_{i,j}$ denotes the current number of times player $i$ has beaten $j$, then the conditional density of $V_1, \ldots, V_n$ is

$$f(x_1, \ldots, x_n) = Ce^{-\sum_i x_i} \prod_{i \neq j} \left( \frac{x_i}{x_i + x_j} \right)^{w_{i,j}} \tag{3.1}$$

for a normalization factor $C$. Our algorithmic approach for strong regret 2 is to simulate $\mathbf{V^{(1)}} = (V_1^{(1)}, \ldots, V_n^{(1)})$ and $\mathbf{V^{(2)}} = (V_1^{(2)}, \ldots, V_n^{(2)})$ independently according to Eq. (3.1), then let

$$I = \text{argmax}_i V_i^{(1)}, \quad J = \text{argmax}_i V_i^{(2)}$$

and choose $I$ and $J$ to play with each other in the next round. (Note that if $I = J$ then (3.1) need not be updated.) For strong regret 1, we simulate only $\mathbf{V^{(1)}} = (V_1^{(1)}, \ldots, V_N^{(1)})$ and choose the two indices with the largest values to play the next game.

However, because directly simulating $\mathbf{V}$ from Eq. (3.1) does not seem computationally feasible (for one thing $C$ is difficult to compute), we utilize the Hasting–Metropolis algorithm to generate a Markov chain whose limiting distribution is given by Eq. (3.1). The Markov chain is defined as follows. When its current state is $\mathbf{x} = (x_1, \ldots, x_n)$, a coordinate that is equally like to be any of $1, \ldots, n$ is selected. If $i$ is selected, a random variable $Y$ is generated from an exponential distribution with mean $x_i$, and if $Y = y$, then $(x_1, \ldots, x_{i-1}, y, x_{i+1}, \ldots, x_n)$ is considered as the candidate next state. In other words, letting $\mathbf{y} = (x_1, \ldots, x_{i-1}, y, x_{i+1}, \ldots, x_n)$, the density function for the candidate next state is

$$q(\mathbf{y} \,|\, \mathbf{x}) = \frac{1}{n} \frac{1}{x_i} e^{-y/x_i}$$

If $\mathbf{x}$ is the current state and $\mathbf{y}$ the candidate next state, then the next state of the Markov chain is $\mathbf{x}$ with probability $\alpha(\mathbf{x}, \mathbf{y})$, or $\mathbf{y}$ with probability $1 - \alpha(\mathbf{x}, \mathbf{y})$, where

$$\alpha(\mathbf{x}, \mathbf{y}) = \min \left\{ \frac{f(\mathbf{y})}{f(\mathbf{x})} \frac{q(\mathbf{x} \,|\, \mathbf{y})}{q(\mathbf{y} \,|\, \mathbf{x})}, 1 \right\}$$

- For strong regret 1, the simulation of the Markov chain stops after, say, $k$ iterations and we use the indices of two largest values of the final state vector as the choice of players for the next round.
- For strong regret 2, we let the simulation of Markov chain stop after $2k$ iterations. We choose index of largest value of the vector at iteration $k$ and $2k$ as the choice of the first player and the second player, respectively.

Suppose that the final stage vector ($k$ iterations for strong regret 1 and $2k$ iterations for strong regret 2) is $x_1^0, \ldots, x_n^0$. Once that round has been completed, and we have updated the values of $w_{i,j}$, we let the initial value of the Markov chain used to obtain the next pair of duelists be $x_1^0, \ldots, x_n^0$. Because the conditional density should not change by much after a single game, we expect this will speed the convergence of the chain. In practice, it turns out that $k = O(N)$ would be enough for each simulation. In addition, we compute $\alpha(\mathbf{x}, \mathbf{y})$, by using the identity $\alpha = \exp(\log(\alpha))$.

https://doi.org/10.1017/S0269964820000601 Published online by Cambridge University Press

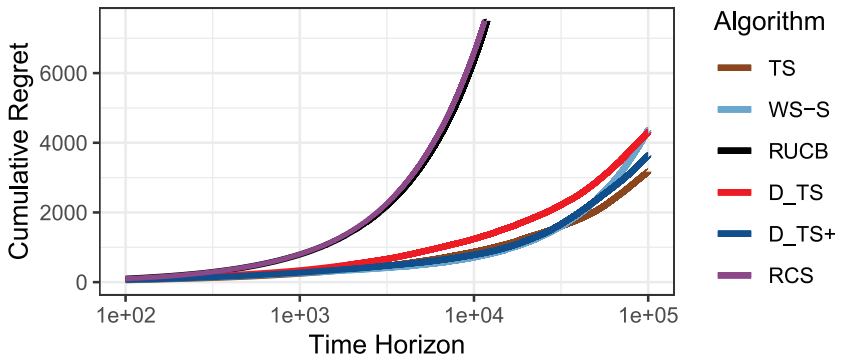


FIGURE **3.** Experiments with five arms on exponential (1) strengths. Replication: 200 times.

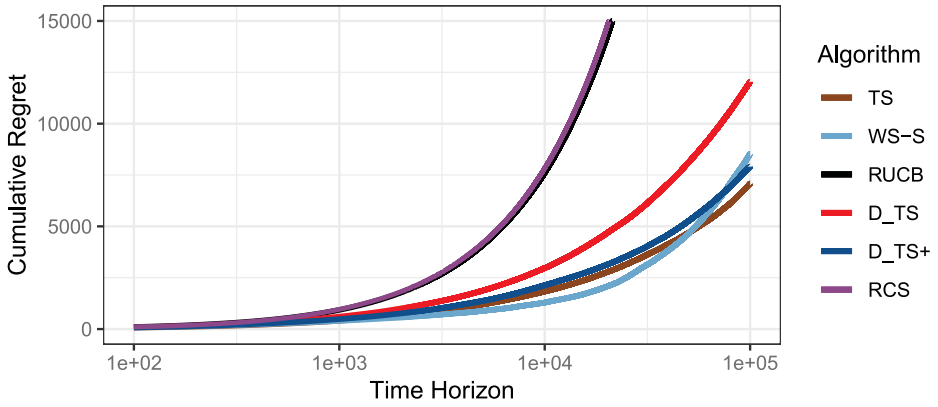


FIGURE **4.** Experiments with five arms on uniform (0, 1) strengths. Replication: 200 times.

### 3.2. Numerical Experiments

We empirically compare our Thompson sampling approach with five benchmarks (WS-S, RUCB, D-TS, D-TS* and RCB) over the simulated data under strong regret 2 criteria. The comparison is conducted in two scenarios, where i.i.d. exponential (1) and uniform (0,1) random variables are generated as the strengths of arms. When strengths are generated from exponential (1), Figure 3 shows that our TS approach empirically seems to outperform all benchmarks, except for some time horizons where WS-S performs only slightly better than TS. When strengths are generated from uniform (0,1), as shown in Figure 4, TS outperforms all benchmarks, except for some time horizons where DT_S+ and WS-S and WS-S perform slightly better than TS. Overall, TS performs either better or in some limited cases only slightly worse than other benchmarks. We were comfortable drawing these conclusions from a possibly small sample size of 200 replications as with this sample size the resulting variation was small enough. We are confident in the ordering of the performance of the approaches and our overall qualitative conclusions.

https://doi.org/10.1017/S0269964820000601 Published online by Cambridge University Press

*ACKNOWLEDGMENTS*

This material is based upon work supported by, or in part by the National Science Foundation under contract/grant number CMMI1662442.

*References*

1. Busa-Fekete, R., Szorenyi, B., Cheng, W., Weng, P., & Hüllermeier, E. (2013). Top-k selection based on adaptive sampling of noisy preferences. *Proceedings of the 30th International Conference on Machine Learning*, in PMLR 28(3), 1094–1102.
2. Chen, B. & Frazier, P.I. (2017). Dueling bandits with weak regret. Preprint arXiv:1706.04304.
3. Komiyama, J., Honda, J., Kashima, H., & Nakagawa, H. (2015). Regret lower bound and optimal algorithm in dueling bandit problem. *Proceedings of The 28th Conference on Learning Theory*, in PMLR 40, 1141–1154.
4. Komiyama, J., Honda, J., & Nakagawa, H. (2016). Copeland dueling bandit problem: Regret lower bound, optimal algorithm, and computationally efficient algorithm. Preprint arXiv:1605.01677.
5. Thompson, W.R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* 25(3/4), 285–294.
6. Urvoy, T., Clerot, F., Féraud, R., & Naamane, S. (2013). Generic exploration and k-armed voting bandits. *Proceedings of the 30th International Conference on Machine Learning*, in PMLR 28(2), 91–99.
7. Wu, H. & Liu, X. (2016). Double Thompson sampling for dueling bandits. *Advances in Neural Information Processing Systems* 29, 649–657.
8. Yue, Y. & Joachims, T. (2009). Interactively optimizing information retrieval systems as a dueling bandits problem. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 1201–1208.
9. Yue, Y. & Joachims, T. (2011). Beat the mean bandit. *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pp. 241–248.
10. Zoghi, M., Whiteson, S., Munos, R., & Rijke, M. (2014). Relative upper confidence bound for the k-armed dueling bandit problem. *Proceedings of the 31st International Conference on Machine Learning*, in PMLR 32(2), 10–18.
11. Zoghi, M., Whiteson, S.A., de Rijke, M., & Munos, R. (2014). Relative confidence sampling for efficient on-line ranker evaluation. *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, pp. 73–82.
12. Zoghi, M., Whiteson, S., de Rijke, M. (2015). Mergerucb: A method for large-scale online ranker evaluation. *Proceedings of the 8th ACM International Conference on Web Search and Data Mining*, pp. 17–26.

## APPENDIX

Proof of Lemma 2.3: If the best arm is the winner of the round $k-1$, the best arm needs to beat all components to win round $k$ no matter in what order, which gives

$$\begin{aligned} P(W_k \mid W_{k-1}) &= \prod_{i=1}^{N} P(\text{best player wins iteration } i) \\ &\geq \prod_{i=1}^{N} \frac{1-(q/p)^{N(k-1)+k+i-1}}{1-(q/p)^{N(k-1)+k+i}} \\ &= \frac{1-(q/p)^{Nk-N+k}}{1-(q/p)^{N(k+1)}} \end{aligned}$$

https://doi.org/10.1017/S0269964820000601 Published online by Cambridge University Press

On the other hand, if the best arm is the challenger at round $k$ and suppose that it comes to play at iteration $j$, $1 \le j \le N$,

$$P(W_k \mid L_{k-1}) = \frac{1-(q/p)}{1-(q/p)^{N(k-1)+k+j}} \prod_{i=j+1}^{N} \frac{1-(q/p)^{N(k-1)+k+i-1}}{1-(q/p)^{N(k-1)+k+i}}$$
$$= \frac{1-(q/p)}{1-(q/p)^{N(k+1)}}$$

■

Proof of Lemma 2.4: Let $r = q/p$. By conditioning on whether the best player wins round $k-1$

$$\begin{aligned} P(L_k) &= P(L_k|W_{k-1})P(W_{k-1}) + P(L_k|L_{k-1})P(L_{k-1}) \\ &= (1 - P(W_k|W_{k-1}))P(W_{k-1}) + (1 - P(W_k|L_{k-1}))P(L_{k-1}) \\ &= \frac{r^{Nk-N+k} - r^{N(k+1)}}{1 - r^{N(k+1)}}(1 - P(L_{k-1})) + \left(\frac{r - r^{N(k+1)}}{1 - r^{N(k+1)}}\right) P(L_{k-1}) \\ &= \frac{r - r^{Nk-N+k}}{1 - r^{N(k+1)}} P(L_{k-1}) + \frac{r^{Nk-N+k} - r^{N(k+1)}}{1 - r^{N(k+1)}} \\ &< rP(L_{k-1}) + r^{N(k-1)+k} \end{aligned}$$

Solving the recursion formula gives us

$$E_p[X] = \frac{n}{2p-1}\left(\frac{1-(q/p)}{1-(q/p)^n} - \frac{1}{n}\right) \quad p \neq \frac{1}{2}$$
$$\mathbb{E}_p[X] = n - 1 \quad p = \frac{1}{2}$$

■

Proof of Lemma 2.5:

$$E_p[X] = \frac{n}{2p-1}\left(\frac{1-(q/p)}{1-(q/p)^n} - \frac{1}{n}\right) \quad p \neq \frac{1}{2}$$
$$\mathbb{E}_p[X] = n - 1 \quad p = \frac{1}{2}$$

Let $r = q/p$ and thus $p = 1/(1+r)$. When $p \neq 1/2$, substitute $p$ by $r$

$$E_p[X] = \left(\frac{n(1-r)}{1-r^n} - 1\right)\frac{1+r}{1-r}$$

For $r > 1$

$$\begin{aligned} \mathbb{E}_p[X] - 2n &= \left(-n + \frac{r^n - 1}{r-1} - 2n\frac{r^n - 1}{1+r}\right)\frac{1+r}{r^n - 1} \\ &= \left(-n - \frac{(2r-2)n}{r+1} + 1\right)\sum_{i=0}^{i-1} r^i \frac{1+r}{rn-1} \\ &< 0 \end{aligned}$$

https://doi.org/10.1017/S0269964820000601 Published online by Cambridge University Press

For $r < 1$

$$\begin{aligned}
\mathbb{E}_p[X] - 2n &= \left( n - \frac{r^n - 1}{r - 1} + 2n \frac{r^n - 1}{1 + r} \right) \frac{1 + r}{1 - r^n} \\
&> \left( n - \frac{r^n - 1}{r - 1} + n(r^n - 1) \right) \frac{1 + r}{1 - r^n} \\
&= \left( - \sum_{i=0}^{n-1} r^i + nr^n \right) \frac{1 + r}{1 - r^n} \\
&< 0
\end{aligned}$$

The proof is complete. ■

https://doi.org/10.1017/S0269964820000601 Published online by Cambridge University Press