Dueling Bandits with Weak Regret

Bangrui Chen [1] **Peter I. Frazier** [1]

Abstract

We consider online content recommendation with implicit feedback through pairwise comparisons, formalized as the so-called dueling bandit problem. We study the dueling bandit problem in the Condorcet winner setting, and consider two notions of regret: the more well-studied strong regret, which is 0 only when both arms pulled are the Condorcet winner; and the less well-studied weak regret, which is 0 if either arm pulled is the Condorcet winner. We propose a new algorithm for this problem, *Winner Stays* (WS), with variations for each kind of regret: WS for weak regret (WS-W) has expected cumulative weak regret that is $O(N^2)$, and $O(N\log(N))$ if arms have a total order; WS for strong regret (WS-S) has expected cumulative strong regret of $O(N^2 + N\log(T))$, and $O(N\log(N) + N\log(T))$ if arms have a total order. WS-W is the first dueling bandit algorithm with weak regret that is constant in time. WS is simple to compute, even for problems with many arms, and we demonstrate through numerical experiments on simulated and real data that WS has significantly smaller regret than existing algorithms in both the weak- and strong-regret settings.

1. Introduction

We consider bandit learning in personalized content recommendation with implicit pairwise comparisons. We offer pairs of items to a user and record implicit feedback on which offered item is preferred, seeking to learn the user's preferences over items quickly, while also ensuring that the fraction of time we fail to offer a high-quality item is small. Implicit pairwise comparisons avoid the inaccuracy of user ratings (Joachims et al., 2007) and the difficulty of engaging users in providing explicit feedback.

We study a model for this setting called the dueling bandit problem (Yue & Joachims, 2009). The items we may offer to the user are called "arms", and we learn about these arms through a sequence of "duels". In each duel, we "pull" two arms and receive noisy feedback from the user telling us which arm is preferred. When an arm is preferred within a duel, we say that the arm has "won the duel".

We study this problem in the Condorcet winner setting, in which we assume the existence of an arm (the Condorcet winner) that wins with probability at least $\frac{1}{2}$ when paired with any of the other arms. In these settings, we consider two notions of regret: "weak regret", in which we avoid regret by selecting the Condorcet winner as either arm in the duel; and "strong-regret", in which we can only avoid regret by setting both arms in the duel to the Condorcet winner.

Weak regret was proposed by Yue et al. (2012) and arises in content recommendation when arms correspond to items, and the user incurs no regret whenever his most preferred item is made available. Examples include in-app restaurant recommendations provided by food delivery services like Grubhub and UberEATS, in which implicit feedback may be inferred from selections, and the user only incurs regret if her most preferred restaurant is not recommended. Examples also include recommendation of online broadcasters on platforms such as Twitch, in which implicit feedback may again be inferred from selections, and the user is fully satisfied as long as her favored broadcaster is listed. Despite its applicability, Yue et al. (2012) is the only paper of which we are aware that studies weak regret, and it does not provide algorithms specifically designed for this setting.

Strong regret has been more widely studied, as discussed below, and has application to choosing ranking algorithms for search (Hofmann et al., 2013). To perform a duel, query results from two rankers are interleaved (Radlinski et al., 2008), and the ranking algorithm that provided the first result chosen by the user is declared the winner of the duel. Strong regret is appropriate in this setting because the user's experience is enhanced by pulling the best arm twice, so that all of that ranker's results are shown.

Our contribution is a new algorithm, *Winner Stays* (WS), with variants designed for the weak (WS-W) and strong regret (WS-S) settings. We prove that WS-W has expected cumulative weak regret that is constant in time, with de-

[1] Cornell University, Ithaca, NY. Correspondence to: Peter I. Frazier <pf98@cornell.edu>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017. Copyright 2017 by the author(s).

pendence on the number of arms N given by $O(N^2)$. If the arms have a total order, we show a tighter bound of $O(N \log N)$. We then prove that WS-S has expected cumulative strong regret that is $O(N^2 + N \log(T))$, and prove that a tighter bound of $O(N \log(N) + N \log(T))$ holds if arms have a total order. These regret bounds are optimal in T, and for weak regret are strictly better than those for any previously proposed algorithm, although at the same time both strong and weak regret bounds are sensitive to the minimum gap in winning probability between arms. We demonstrate through numerical experiments on simulated and real data that WS-W and WS-S significantly outperform existing algorithms on strong and weak regret.

The paper is structured as follows. Section 2 reviews related work. Section 3 formulates our problem. Section 4 introduces the *Winner Stays* (WS) algorithm: Section 4.1 defines WS-W for the weak regret setting; Section 4.2 proves that WS-W has cumulative expected regret that is constant in time; Section 4.3 defines WS-S for the strong regret setting and bounds its regret. Section 4.4 disusses a simple extension of our theoretical results to the utility-based bandit setting, which is used in our numerical experiments. Section 5 compares WS with three benchmark algorithms using both simulated and real datasets, finding that WS outperforms these benchmarks on the problems considered.

2. Related Work

Most work on dueling bandits focuses on strong regret. Yue et al. (2012) shows that the worst-case expected cumulative strong regret up to time T for any algorithm is $\Omega(N \log(T))$. Algorithms have been proposed that reach this lower bound under the Condorcet winner assumption in the finite-horizon setting: Interleaved Filter (IF) (Yue et al., 2012) and Beat the Mean (BTM) (Yue & Joachims, 2011). Relative Upper Confidence Bound (RUCB) (Zoghi et al., 2014) also reaches this lower bound in the horizonless setting. Relative Minimum Empirical Divergence (RMED) (Komiyama et al., 2015) is the first algorithm to have a regret bound that matches this lower bound. Zoghi et al. (2015) proposed two algorithms, Copeland Confidence Bound (CCB) and Scalable Copeland Bandits (SCB), which achieve an optimal regret bound without assuming existence of a Condorcet winner.

While weak regret was proposed in Yue et al. (2012), it has not been widely studied to our knowledge, and despite its applicability we are unaware of papers that provide algorithms designed for it specifically. While one can apply algorithms designed for the strong regret setting to weak regret, and use the fact that strong dominates weak regret to obtain weak regret bounds of $O(N \log(T))$, these are looser than the constant-in-T bounds that we show.

Active learning using pairwise comparisons is also closely related to our work. Jamieson & Nowak (2011) considers an active learning problem that is similar to our problem in that the primary goal is to sort arms based on the user's preferences, using adaptive pairwise comparisons. It proposes a novel algorithm, the Query Selection Algorithm (QSA), that uses an expected number of operations of $d \log(N)$ to sort N arms, where d is the dimension of the space in which the arms are embedded, rather than $N \log(N)$. Busa-Fekete et al. (2013) and Busa-Fekete et al. (2014) consider top-k element selection using adaptive pairwise comparisons. They propose a generalized racing algorithm focusing on minimizing sample complexity. (Pallone et al., 2017) studies adaptive preference learning across arms using pairwise preferences. They show that a greedy algorithm is Bayes-optimal for an entropy objective. While similar in that they use pairwise comparisons, these algorithms are different in focus from the current work because they do not consider cumulative regret.

3. Problem Formulation

We consider N items (arms). At each time $t = 1, 2, \ldots$, the system chooses two items and shows them to the user, i.e., the system performs a duel between two arms. The user then provides binary feedback indicating her preferred item, determining which arm wins the duel. This binary feedback is random, and is conditionally independent of all past interactions given the pair of arms shown. We let $p_{i,j}$ denote the probability that the user gives feedback indicating a preference for arm i, when shown arms i and j. If the user prefers arm i over arm j, we assume $p_{i,j} > 0.5$. We also assume symmetry: $p_{i,j} = 1 - p_{j,i}$.

We assume arm 1 is a Condorcet winner, i.e., that $p_{1,i} > 0.5$ for $i = 2, \cdots, N$. In some results, we also consider the setting in which arms have a total order, by which we mean that the arms are ordered so that $p_{i,j} > 0.5$ for all $i < j$. The total order assumption implies transitivity.

We let $p = \min_{p_{i,j} > 0.5} p_{i,j} > 0.5$ be a lower bound on the probability that the user will choose her favourite arm.

We consider both weak and strong regret in its binary form. The single-period *weak regret* incurred at this time is $r(t) = 1$ if we do not pull the best arm and $r(t) = 0$ otherwise. The single-period *strong regret* is $r(t) = 1$ if we do not pull the best arm twice and $r(t) = 0$ otherwise. We also consider utility-based extensions of weak and strong regret in Section 4.4.

We use the same notation $r(t)$ to denote strong and weak regret, and rely on context to distinguish the two cases. In both cases, we define the cumulative regret up to time T to be $R(T) = \sum_{t=1}^{T} r(t)$. We measure the quality of an algorithm by its expected cumulative regret.

4. Winner Stays

We now propose an algorithm, called *Winner Stays* (WS), with two variants: WS-W designed for weak regret; and WS-S for strong regret. Section 4.1 introduces WS-W and illustrates its dynamics. Section 4.2 proves the expected cumulative weak regret of WS-W is $O(N^2)$ under the Condorcet winner setting, and $O(N\log(N))$ under the total order setting. Section 4.3 introduces WS-S and proves that its expected cumulative strong regret is $O(N^2 + N\log(T))$ under the Condorcet winner setting, and $O(N\log(T) + N\log(N))$ under the total order setting, both of which have optimal dependence on T. Section 4.4 extends our theoretical results to utility-based bandits.

4.1. Winner Stays with Weak Regret (WS-W)

We now present WS-W, first defining some notation. Let $q_{i,j}(t)$ be the number of times that arm i has defeated arm j in a duel, up to and including time t. Then, define $C(t,i) = \sum_{j\neq i} q_{i,j}(t) - q_{j,i}(t)$. $C(t,i)$ is the difference between the number of duels won and lost by arm i, up to time t. With this notation, we define WS-W in Algorithm 1.

Algorithm 1 WS-W

```
Input: arms $1, \cdots, N$
for $t = 1, 2, \cdots$ do
    Step 1: Pick $i_t = \arg\max_i C(t-1,i)$, breaking ties as follows:
      • If $t > 1$ and $i_{t-1} \in \arg\max_i C(t-1,i)$, set $i_t = i_{t-1}$.
      • Else if $t > 1$ and $j_{t-1} \in \arg\max_i C(t-1,i)$, set $i_t = j_{t-1}$.
      • Else choose $i_t$ uniformly at random from $\arg\max_i C(t-1,i)$.
    Step 2: Pick $j_t = \arg\max_{j\neq i_t} C(t-1,j)$, breaking ties as follows:
      • If $t > 1$ and $i_{t-1} \in \arg\max_{i\neq i_t} C(t-1,i) \setminus \{i_t\}$, set $j_t = i_{t-1}$.
      • Else if $t > 1$ and $j_{t-1} \in \arg\max_{i\neq i_t} C(t-1,i) \setminus \{i_t\}$, set $j_t = j_{t-1}$.
      • Else choose $j_t$ uniformly at random from $\arg\max_j C(t-1,j) \setminus \{i_t\}$.
    Step 3: Pull arms $i_t$ and $j_t$;
    Step 4: Observe noisy binary feedback and update $C(t,i_t)$ and $C(t,j_t)$;
end
```

WS-W's pulls can be organized into *iterations*, each of which consists of a sequence of pulls of the same pair of arms, and *rounds*, each of which consists of a sequence of iterations in which arms that lose an iteration are not visited again until the next round. We first describe iterations and rounds informally with an example and in Figure 1 before presenting our formal analysis.

Example: At time $t = 1$, $C(0,i) = 0$ for all i, and WS-W pulls two randomly chosen arms. Suppose it pulls arms $i_1 = 1$, $j_1 = 2$ and arm 1 wins. Then $C(1,i)$ is 1 for arm 1, -1 for arm 2, and 0 for the other arms. This first pull is an iteration of length 1, arm 1 is the winner, and arm 2 is the loser. This iteration is in the first round. We call $t_1 = 1$ the start of the first round, and $t_{1,1} = 1$ the start of the first iteration in the first round.

At time $t = 2$, $C(t-1,i)$ is largest for arm 1 so WS-W chooses $i_2 = 1$. Since $C(t-1,i)$ is -1 for arm 2 and 0 for the other arms, WS-W chooses j_2 at random from arms 3 through N (suppose $N > 2$). Suppose it chooses arm $j_2 = 3$. This pair of arms (1 and 3) is different from the pair pulled in the previous iteration (1 and 2), so $t_{1,2} = 2$ is the start of the second iteration (in the first round).

WS-W continues pulling arms 1 and 3 until $C(t,i)$ is -1 for one of these arms and 2 for the other. WS-W continues to pull only arms 1 and 3 until one has $C(t,i) = 2$ even though this may involve times when $C(t,i)$ is 0 for both arms 1 and 3, causing them to be tied with arms 4 and above, because we break ties to prioritize pulling previously pulled arms. The sequence of times when we pull arms 1 and 3 is the second iteration. The arm that ends the iteration with $C(t,i) = 2$ is the winner of that iteration.

WS-W continues this process, performing $N-1$ iterations on different pairs of arms, pitting the winner of each iteration against a previously unplayed arm in the next iteration. This sequence of iterations is the first round. The winner of the final iteration in the first round, call it arm $Z(1)$, has $C(t,Z(1)) = N-1$ and all other arms $j \neq Z(1)$ have $C(t,j) = -1$.

The second round begins on the next pull after the end of the first round, at time t_2. WS-W again performs $N-1$ iterations, playing $Z(1)$ in the first iteration. Each iteration has a winner that passes to the next iteration.

WS-W repeats this process for an infinite number of rounds. Each round is a sequence of $N-1$ iterations, and an arm that loses an iteration is not revisited until the next round. Figure 1 illustrates these dynamics, and we formalize the definition of round and iteration in the next section.

4.2. Analysis of WS-W

In this section, we analyze the weak regret of WS-W. After presenting definitions and preliminary results, we prove WS-W has expected cumulative weak regret bounded by $O(N\log(N))$ when arms have a total order. Then, in the more general Condorcet winner setting, we prove WS-W has expected cumulative weak regret bounded by $O(N^2)$. We leave the proofs of all lemmas to the supplement.

We define t_ℓ, the *beginning of round* ℓ, and $Z(\ell-1)$, the

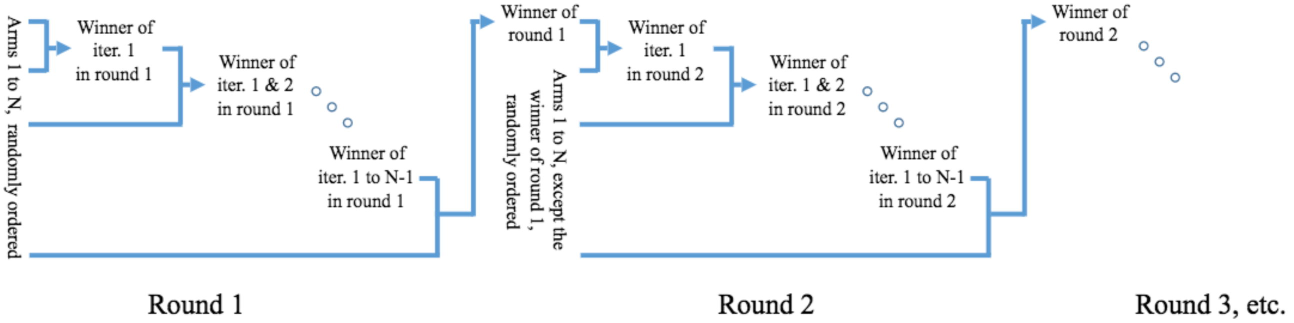

Arms 1 to N, randomly ordered
Winner of iter. 1 in round 1
Winner of iter. 1 & 2 in round 1
Winner of iter. 1 to N-1 in round 1
Winner of round 1
Arms 1 to N, except the winner of round 1, randomly ordered
Winner of iter. 1 in round 2
Winner of iter. 1 & 2 in round 2
Winner of iter. 1 to N-1 in round 2
Winner of round 2
Round 1
Round 2
Round 3, etc.

Figure 1. Our analysis of WS-W decomposes its behavior into a sequence of rounds. In each round, pairs of arms play each other in a sequence of iterations. The winner from an iteration passes on to play a new arm in the next iteration randomly selected from those that have not yet played in the round. At the end of a round, the round's winner is considered first in the next round.

winner of round ℓ, as the unique time and arm such $C(t_\ell - 1, Z(\ell-1)) = (N-1)(\ell-1)$ and $C(t_\ell - 1, i) = -\ell + 1$ for all $i \neq Z(\ell-1)$.

We define $t_{\ell,k}$, the *beginning of iteration k in round ℓ*, as the first time we pull the k^{th} unique pair of arms in the ℓ^{th} round. We let $T_{\ell,k}$ be the number of successive pulls of this pair of arms.

We additionally define terminology to describe arms pulled in an iteration. In a duel between arms i and j with $p_{i,j} > 0.5$, arm i is called the *better arm* and arm j is called the *worse arm*. We say that an arm i is the *incumbent* in iteration k iteration and round ℓ if $C(t_{\ell,k} - 1, i) > 0$. A unique such arm exists except when $\ell = k = 1$. When $\ell = k = 1$, the incumbent is the better of the two arms being played. We call the arm being played that is not the incumbent the *challenger*.

Using these definitions, we present our first pair of results toward bounding the expected cumulative weak regret of WS-W. They bound the number of pulls in an iteration.

Lemma 1. The conditional expected length of iteration k in round ℓ, given the arms being pulled, is bounded above by $\frac{N(\ell-1)+k}{2p-1}$ if the incumbent is worse than the challenger, and by $\frac{1}{2p-1}$ if the incumbent is better than the challenger.

Lemma 1 shows that iterations with a worse incumbent use more pulls. We then bound the number of iterations with a worse incumbent.

Lemma 2. Under the total order assumption, the conditional expected number of future iterations with an incumbent worse than the challenger, given history up to time $t_{\ell,k}$, is bounded above by $\frac{2p^2}{(2p-1)^3}(\log(N)+1)$ for any $k, \ell \geq 1$.

Lemma 2 implies that the incumbent is worse than the challenger in finitely many iterations with probability 1. We now bound the tail distribution of the last such round.

Lemma 3. Let L denote the smallest ℓ such that no round $\ell' > \ell$ contains an iteration in which the incumbent is worse than the challenger. Then $P(L \geq \ell) \leq \left(\frac{1-p}{p}\right)^{\ell}$.

To present our final set of preliminary lemmas, we define several indicator functions. Let $B(\ell, k)$ be 1 when the incumbent in iteration k of round ℓ is better than the challenger. Let $D(\ell)$ be 1 if arm 1 (the best arm) is the incumbent at the beginning of iteration 1 of round ℓ. Denote $\bar{B}(\ell,k) = 1 - B(\ell,k)$ and $\bar{D}(\ell) = 1 - D(\ell)$. Let $V(\ell,k)$ be 1 if $D(\ell) = 1$ and arm 1 loses in any iteration 1 through $k-1$ of round ℓ.

We may only incur weak regret during round ℓ iteration k if $\bar{D}(\ell) = 1$, or if $V(\ell, k') = 1$ for some $k' < k$. We will separately bound the regret incurred in these two different scenarios. Moreover, our bound on the number of pulls, and thus the regret incurred, in this iteration will depend on whether $B(\ell,k) = 1$ or $\bar{B}(\ell,k) = 1$. This leads us to state four inequalities in the following pair of lemmas, which we will in turn use to show Theorem 1. The first lemma applies in both the total order and Condorcet settings, while the second applies only in the total order setting. When proving Theorem 2 we replace Lemma 5 by an alternate pair of inequalities.

Lemma 4.

$$\mathbb{E}[\bar{D}(\ell)B(\ell,k)T_{\ell,k}] \leq \frac{1}{2p-1}\left(\frac{1-p}{p}\right)^{\ell-1},$$

$$\mathbb{E}[V(\ell,k)B(\ell,k)T_{\ell,k}] \leq \frac{1}{2p-1}\left(\frac{1-p}{p}\right)^{\ell}.$$

Lemma 5. Under the total order assumption:

- $\mathbb{E}\left[\sum_{k=1}^{N-1} \bar{D}(\ell)\bar{B}(\ell,k)T_{\ell,k}\right]$ is bounded above by $\left(\frac{1-p}{p}\right)^{\ell-1} \frac{2N\ell p^2}{(2p-1)^4}(\log(N)+1)$.

- $\mathbb{E}\left[\sum_{k=1}^{N-1} V(\ell,k)\bar{B}(\ell,k)T_{\ell,k}\right]$ is bounded above by $\left(\frac{1-p}{p}\right)^{\ell} \frac{2N\ell p^2}{(2p-1)^4}(\log(N)+1)$.

We now state our main result for the total order setting, which shows that the expected cumulative weak regret is $O\left(\frac{N\log(N)}{(2p-1)^5}\right)$.

Theorem 1. The expected cumulative weak regret of WS-W is bounded by $\left[\frac{2p^3}{(2p-1)^6}N(\log(N)+1)+\frac{N}{(2p-1)^2}\right]$ under the total order assumption.

Proof. Iterations can be divided into two types: those in which the incumbent is better than the challenger, and those where the incumbent is worse.

We first bound expected total weak regret incurred in the first type of iteration, and then below bound that incurred in the second type. In this first bound, observe that we incur weak regret during round ℓ if $D(\ell)=0$, or if $D(\ell)=1$ but arm 1 loses to some other arm during this round. Under the second scenario, we do not incur any regret until arm 1 loses to another arm.

Thus, the expected weak regret incurred during iterations with a better incumbent is bounded by

$$\mathbb{E}\left[\sum_{\ell=1}^{\infty}\sum_{k=1}^{N-1} B(\ell,k)T_{\ell,k}\bar{D}(\ell)+\sum_{\ell=1}^{\infty}\sum_{k=1}^{N-1} B(\ell,k)T_{\ell,k}V(\ell,k)\right].$$

The first part of this summation can be bounded by the first inequality in Lemma 4 to obtain

$$\begin{aligned}&\mathbb{E}\left[\sum_{\ell=1}^{\infty}\sum_{k=1}^{N-1} B(\ell,k)T_{\ell,k}\bar{D}(\ell)\right]\\ &\leq \sum_{\ell=1}^{\infty}\left(\frac{1-p}{p}\right)^{\ell-1}\frac{N}{2p-1}=\frac{pN}{(2p-1)^2}.\end{aligned}$$

The second part of this summation can be bounded by the second inequality in Lemma 4 to obtain

$$\begin{aligned}&\mathbb{E}\left[\sum_{\ell=1}^{\infty}\sum_{k=1}^{N-1} B(\ell,k)T_{\ell,k}V(\ell,k)\right]\\ &\leq \sum_{\ell=1}^{\infty}\frac{N}{2p-1}\left(\frac{1-p}{p}\right)^{\ell}=\frac{N(1-p)}{(2p-1)^2}.\end{aligned}$$

Thus, the cumulative expected weak regret incurred during iterations with a better incumbent is bounded by $\frac{N}{(2p-1)^2}$.

Now we bound the expected weak regret incurred during iterations where the incumbent is worse than the challenger. This is bounded by

$$\mathbb{E}\left[\sum_{\ell=1}^{\infty}\sum_{k=1}^{N-1} \bar{B}(\ell,k)T_{\ell,k}\bar{D}(\ell)+\sum_{\ell=1}^{\infty}\sum_{k=1}^{N-1} \bar{B}(\ell,k)T_{\ell,k}V(\ell,k)\right].$$

The first term in the summation can be bounded by the first inequality of Lemma 5 to obtain

$$\begin{aligned}&\mathbb{E}\left[\sum_{\ell=1}^{\infty}\sum_{k=1}^{N-1} \bar{B}(\ell,k)T_{\ell,k}\bar{D}(\ell)\right]\\ &\leq \sum_{\ell=1}^{\infty}\frac{2N\ell p(1-p)}{(2p-1)^4}(\log(N)+1)\left(\frac{1-p}{p}\right)^{\ell-1}\\ &=\frac{2Np^4}{(2p-1)^6}(\log(N)+1).\end{aligned}$$

The second term in the summation can be bounded by the first inequality of Lemma 5 to obtain

$$\begin{aligned}&\mathbb{E}\left[\sum_{\ell=1}^{\infty}\sum_{k=1}^{N-1} \bar{B}(\ell,k)T_{\ell,k}V(\ell,k)\right]\\ &\leq \sum_{\ell=1}^{\infty}\frac{2N\ell p^2}{(2p-1)^4}(\log(N)+1)\left(\frac{1-p}{p}\right)^{\ell}\\ &=\frac{2p^3(1-p)}{(2p-1)^6}N(\log(N)+1).\end{aligned}$$

Thus, the cumulative expected weak regret incurred during iterations with a worse incumbent is bounded by $\frac{2p^3}{(2p-1)^6}N(\log(N)+1)$.

Summing these two bounds, the cumulative expected weak regret is bounded by $\left[\frac{2p^3}{(2p-1)^6}N(\log(N)+1)+\frac{N}{(2p-1)^2}\right]$. □

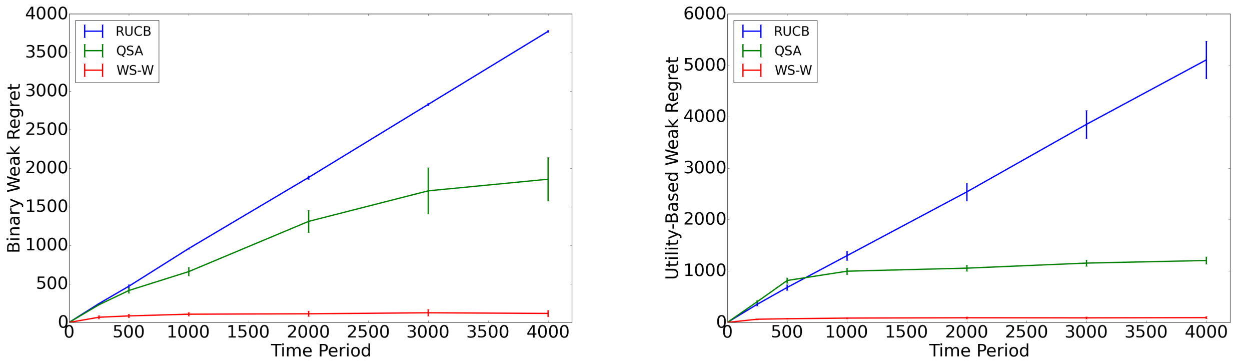
We prove the following result for the Condorcet winner setting in a similar manner in the supplement.

Theorem 2. The expected cumulative weak regret of WS-W is bounded by $\frac{N}{(2p-1)^2}+\frac{pN^2}{(2p-1)^3}$ under the Condorcet winner setting.

4.3. Winner Stays with Strong Regret (WS-S)

In this section, we define a version of WS for strong regret, WS-S, which uses WS-W as a subroutine. WS-S is defined in Algorithm 2

Each round of WS-S consists of an exploration phase and an exploitation phase. The length of the exploitation phase increases exponentially with the number of phases. Changing the parameter β balances the lengths of these phases, and thus balances between exploration and exploitation. Our theoretical results below guide choosing β.



(a) Simulated Dataset

(b) Yelp Academic Dataset

Figure 2. Comparison of the weak regret between WS-W, RUCB and QSA using simulated data, and the Yelp academic dataset. In both experiments, WS-W outperforms RUCB and QSA, provided constant expected cumulative weak regret.

Algorithm 2 WS-S

```
Input: $\beta > 1$, arms $1, \cdots, N$
for $\ell = 1, 2, \cdots$ do
    Exploration phase: Run the $\ell^{th}$ round of WS-W.
    Exploitation phase: Let $Z(\ell)$ be the index of the best arm at the end of the $\ell^{th}$ round. For the next $\lfloor \beta^\ell \rfloor$ time periods, pull arms $Z(\ell)$ and $Z(\ell)$ and ignore the feedback.
end
```

We now bound the cumulative strong regret of this algorithm under both the total order and Condorcet winner settings:

Theorem 3. If there is a total order among arms, then for $1 < \beta < \frac{p}{1-p}$, the expected cumulative strong regret for WS-S is bounded by $\left[\frac{2p^3}{(2p-1)^6} N(\log(N) + 1) + \frac{N \log_\beta(T(\beta-1))}{2p-1}\right]$.

Proof. Suppose at time T, we are in round ℓ. Then $\beta + \cdots + \beta^\ell \leq T$. Solving for ℓ, we obtain $\ell \leq \log_\beta(T(\beta-1))$.

We bound the expected strong regret up to time T. The expected regret can be divided in two parts: the regret occuring during the exploration phase; and the regret occuring during the exploitation phase.

First we focus on regret incurred during exploration. We never pull the same arm twice during this phase, and so regret is incurred in each time period. To bound regret incurred during exploration, we bound the length of time spent in this phase.

The length of time spent in exploration up to the end of round ℓ with a better incumbent is bounded by $\frac{(N-1)\ell}{2p-1}$. The length of time spent with a worse incumbent, based on the proof of Theorem 1, is bounded by $\frac{2p^3}{(2p-1)^6} N(\log(N) + 1)$.

Now we focus on regret incurred during exploitation. The probability we have identified the wrong arm at the end of the i^{th} round is less than $\left(\frac{1-p}{p}\right)^i$. Thus, the expected regret incurred during this phase up until the end of the ℓ^{th} round is bounded by $\sum_{i=1}^{\ell} \left(\frac{1-p}{p}\right)^i \times \beta^i \leq \ell$.

Overall, this implies that the strong expected regret up to time T (recall that T is in round ℓ) is bounded by

$$\left[\frac{2p^3}{(2p-1)^6} N(\log(N) + 1) + \ell + \frac{(N-1)\ell}{2p-1}\right]$$
$$\leq \left[\frac{2p^3}{(2p-1)^6} N(\log(N) + 1) + \frac{N \log_\beta(T(\beta-1))}{2p-1}\right].$$

Thus, the expected strong regret up to time T is $O(N \log(T) + N \log(N))$. □

Theorem 4. Under the Condorcet winner setting and for $1 < \beta < \frac{p}{1-p}$, the expected cumulative strong regret for WS-S is bounded by $\left[\frac{N^2 p}{(2p-1)^2} + \frac{N \log(T(\beta-1))}{(2p-1)\log(\beta)}\right]$.

Proof. The proof mirrors that of Theorem 3, with the only difference being that we bound the length of exploration with a worse incumbent using the proof of Theorem 2 rather than Theorem 1, and the bound is $O(N^2)$. Due to its similarity, the proof is omitted. □

These results provide guidance on the choice of β. If β is too close to 1, then we spend most of the time in the exploration phase, which is guaranteed to generate strong regret. The last inequality in the proof of Theorem 3 suggests that asymptotic regret will be smallest if we choose β as large as possible without going beyond the $p/(1-p)$ threshold. Indeed, if β is too large, then WS-S may incur large regret in early exploitation stages when we have finished only a few rounds of exploration. In our numerical experiments we set $\beta = 1.1$, which satisfies the $p/(1-p)$ constraint assumed by our theory if $p > \beta/(1+\beta) \approx .524$. With a properly chosen β, the numerical experiments in section 5.2 suggest WS-S performs better than previously devised algorithms. At the same time, the best choice of β is dependent on p. Modifying WS-S to eliminate parameters that must be chosen with knowledge of p is left for future work.

Our regret bound grows as p, which is the minimal gap between two arms, shrinks, and p tends to decrease as the number of arms N increases. Other dueling bandit algorithm for strong regret, such as RUCB and RMED, have regret bounds with better dependence on the gaps between arms. Modifying WS-S to provide improved dependence on these gaps is also left for future work.

4.4. Extension to Utility-Based Regret

We now briefly discuss utility-based extensions of weak and strong regret for the total order setting, following utility-based bandits studied in Ailon et al. (2014). Our regret bounds also apply here, with a small modification.

Suppose that the user has a utility u_i associated with each arm i. Without loss of generality, we assume $u_1 > u_2 > \cdots > u_N$, and as in the total order setting, we require that $p_{i,j} > 0.5$ when $i < j$. Typically the $p_{i,j}$ would come from the utilities of arms i and j via a generative model. We give an example in our numerical experiments.

Then, the single-period *utility-based weak regret* is $r(t) = u_1 - \max\{u_{i_t}, u_{j_t}\}$, which is the difference in utility between the best arm overall and the best arm that the user can choose from those offered. The single-period *utility-based strong regret* is $r(t) = u_1 - \frac{u_{i_t}+u_{j_t}}{2}$. To get zero regret under strong regret, the best arm must be pulled twice.

Our results from Section 4 carry through to this more general regret setting. Let $R = u_1 - u_N$ be the maximum single-period regret. Then, the expected cumulative utility-based weak regret for WS-W is $O\left(R\frac{N\log(N)}{(2p-1)^5}\right)$, and the expected cumulative utility-based strong regret for WS-S is $O(R\left[N\log(T) + N\log(N)\right])$.

5. Numerical Experiments

In this section, we evaluate WS under both the weak and strong regret settings, considering both their original (binary) and utility-based versions. In the weak regret setting, we compare WS-W with RUCB and QSA. In the strong regret setting, we compare WS-S with 7 benchmarks including RUCB and Relative Minimum Empirical Divergence (RMED) by Komiyama et al. (2015). We also include an experiment violating the total order assumption in Section 11 in the supplement. WS outperforms all benchmarks tested in these numerical experiments.

5.1. Weak Regret

We now compare WS-W with QSA and RUCB using simulated data and the Yelp academic dataset (Yelp, 2012).

5.1.1. Simulated Data

In this example, we compare WS-W with RUCB and QSA on a problem with 50 arms and binary weak regret. Each arm is a 20-dimensional vector uniformly generated from the unit circle. We assume $p_{i,j} = 0.8$ for all $i < j$.

The results are summarized in Figure 2a. RUCB has approximately linear regret over the time horizon pictured. This is common in the dueling bandits literature, where many algorithms require $\sim 10^4$ comparisons before they achieve $\log(T)$ cumulative regret for 50 arms. WS-W finds the optimal arm after ~ 500 comparisons and has a regret that is consistent with our theoretically established constant expected cumulative weak regret.

5.1.2. Yelp Academic Dataset

In this example, we compare WS-W with RUCB and QSA using the Yelp academic dataset (Yelp, 2012) and utility-based weak regret.

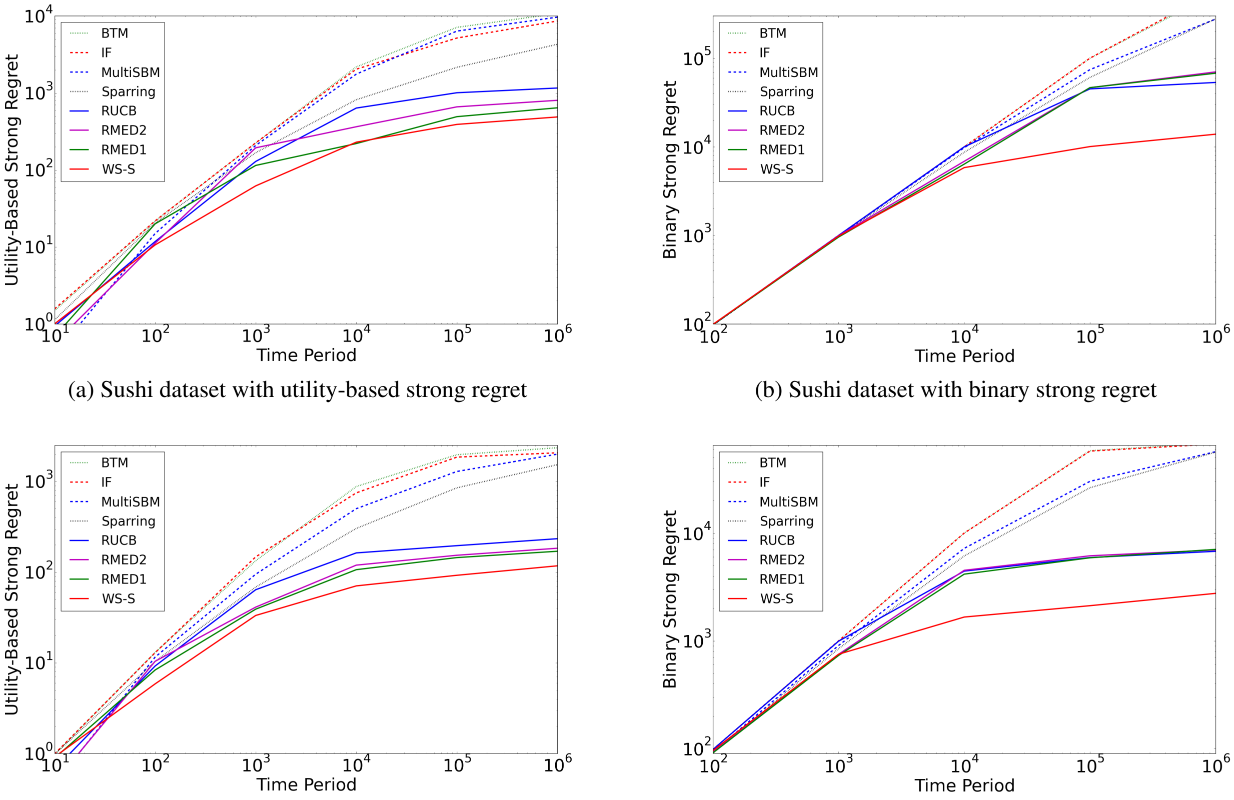
We choose 100 restaurants from Las Vegas as our arms. Associated with each arm (restaurant) i is a 20-dimensional feature vector A_i, calculated using doc2vec (Rehurek & Sojka, 2010) from its reviews. We select 49 users who have reviewed at least 20 of these 100 restaurants. For each user, we model their utility for restaurant i as $u_i = A_i \cdot \theta$, where θ is a 20-dimensional vector of preferences. We infer θ for each user using linear regression.

To model $p_{i,j}$, we then use the probit model. We let $\hat{\sigma}^2$ be the estimated variance of the residuals from the linear regression above. When presented with two restaurants, we model the user as taking independent random draws from a normal distribution with means u_i and u_j respectively and variances $\hat{\sigma^2}$, and choosing the restaurant with the larger draw. This gives $p_{ij} = \Phi(u_i - u_j)$, where $\Phi(\cdot)$ is the cdf for the normal distribution with mean 0 and variance $2\hat{\sigma}^2$.

We simulate performance for each user separately, and then average the results. These results are summarized in Figure 2b. WS-W outperforms RUCB and QSA, finding the optimal restaurant after ~ 500 iterations.

5.2. Strong Regret

In this section, we compare WS-S using binary and utility-based strong regret with 7 benchmarks from the literature. We use the sushi and MSLR datasets, which were previously used by Komiyama et al. (2016) and Zoghi et al. (2015) respectively to evaluate dueling bandit algorithms.



(a) Sushi dataset with utility-based strong regret

(b) Sushi dataset with binary strong regret

(c) MSLR dataset with utility-based strong regret

(d) MSLR dataset with binary strong regret

Figure 3. Comparison of the strong regret between WS-S and 7 benchmarks on the sushi and MSLR datasets. For utility-based strong regret, we start our plot from $t = 10$ since the performance of all algorithms are close to each other before $t = 10$. For the same reason, we start our plot from $t = 100$ for the binary strong regret. WS-S outperforms all benchmarks in all settings studied.

The sushi dataset (Komiyama et al., 2016) contains 16 arms corresponding to types of sushi, with pairwise preferences inferred from data on sushi preferences from 5000 users in Kamishima (2003). The MSLR dataset has 5 arms, corresponding to ranking algorithms, with pairwise preferences provided in Zoghi et al. (2015). We give preference matrices $(p_{i,j})$ for both datasets in the supplement. For utility-based regret, we define $u_i = 2(1 - p_{1,i})$.

WS-S has a user-defined parameter β. In our experiments we set $\beta = 1.1$. The corresponding minimum p for which our theoretical bounds hold is $\beta/(1+\beta) \approx 0.52$. We recommend $\beta \approx 1.1$ for problems of 20 arms or fewer, and β closer to 1 for those problems with more arms that are likely to have p closer to $1/2$. We also conduct a sensitivity analysis of β in the supplement.

Figure 3 shows the results of our comparisons. WS-S outperforms all 7 benchmarks considered on both datasets using both variants of strong regret.

6. Conclusion

In this paper, we consider dueling bandits for online content recommendation using both weak and strong regret.

We propose a new algorithm, WS, with variants designed for the weak regret (WS-W) and strong regret (WS-S) settings. We prove WS has constant weak regret and optimal strong regret in T. In numerical experiments, WS outperforms all benchmarks considered on both simulated and real datasets.

Acknowledgements

The authors were partially supported by NSF CAREER CMMI-1254298, NSF CMMI-1536895, NSF IIS-1247696, NSF DMR-1120296, AFOSR FA9550-12-1-0200, AFOSR FA9550-15-1-0038, and AFOSR FA9550-16-1-0046.

References

Ailon, Nir, Karnin, Zohar Shay, and Joachims, Thorsten. Reducing dueling bandits to cardinal bandits. In *ICML*, volume 32, pp. 856–864, 2014.

Busa-Fekete, Róbert, Szörényi, Balázs, Cheng, Weiwei, Weng, Paul, and Hüllermeier, Eyke. Top-k selection based on adaptive sampling of noisy preferences. In *ICML (3)*, pp. 1094–1102, 2013.

Busa-Fekete, Róbert, Hüllermeier, Eyke, and Szörényi, Balázs. Preference-based rank elicitation using statistical models: The case of mallows. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pp. 1071–1079, 2014.

Hofmann, Katja, Whiteson, Shimon, and Rijke, Maarten De. Fidelity, soundness, and efficiency of interleaved comparison methods. *ACM Transactions on Information Systems (TOIS)*, 31(4):17, 2013.

Jamieson, Kevin G and Nowak, Robert. Active ranking using pairwise comparisons. In *Advances in Neural Information Processing Systems*, pp. 2240–2248, 2011.

Joachims, Thorsten, Granka, Laura, Pan, Bing, Hembrooke, Helene, Radlinski, Filip, and Gay, Geri. Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM Transactions on Information Systems (TOIS)*, 25(2):7, 2007.

Kamishima, Toshihiro. Nantonac collaborative filtering: recommendation based on order responses. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 583–588. ACM, 2003.

Komiyama, Junpei, Honda, Junya, Kashima, Hisashi, and Nakagawa, Hiroshi. Regret lower bound and optimal algorithm in dueling bandit problem. In *COLT*, pp. 1141–1154, 2015.

Komiyama, Junpei, Honda, Junya, and Nakagawa, Hiroshi. Copeland dueling bandit problem: Regret lower bound, optimal algorithm, and computationally efficient algorithm. *arXiv preprint arXiv:1605.01677*, 2016.

Pallone, Stephen N, Frazier, Peter I, and Henderson, Shane G. Bayes-optimal entropy pursuit for active choice-based preference learning. *arXiv preprint arXiv:1702.07694*, 2017.

Radlinski, Filip, Kurup, Madhu, and Joachims, Thorsten. How does clickthrough data reflect retrieval quality? In *Proceedings of the 17th ACM conference on Information and knowledge management*, pp. 43–52. ACM, 2008.

Rehurek, Radim and Sojka, Petr. Software framework for topic modelling with large corpora. In *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Citeseer, 2010.

Yelp, Inc. Yelp academic dataset, 2012. URL `https://www.yelp.com/dataset_challenge`.

Yue, Yisong and Joachims, Thorsten. Interactively optimizing information retrieval systems as a dueling bandits problem. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 1201–1208. ACM, 2009.

Yue, Yisong and Joachims, Thorsten. Beat the mean bandit. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pp. 241–248, 2011.

Yue, Yisong, Broder, Josef, Kleinberg, Robert, and Joachims, Thorsten. The k-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78(5): 1538–1556, 2012.

Zoghi, Masrour, Whiteson, Shimon, Munos, Remi, Rijke, Maarten de, et al. Relative upper confidence bound for the k-armed dueling bandit problem. In *JMLR Workshop and Conference Proceedings*, number 32, pp. 10–18. JMLR, 2014.

Zoghi, Masrour, Karnin, Zohar S, Whiteson, Shimon, and De Rijke, Maarten. Copeland dueling bandits. In *Advances in Neural Information Processing Systems*, pp. 307–315, 2015.