Supplementary Materials: Dueling Bandits with Weak Regret

June 12, 2017

Abstract

This note contains supplementary materials for *Dueling Bandits with Weak Regret*.

0. Gambler's Ruin Lemma

In our analysis of WS-W, we will use results from a special case of the Gambler's ruin problem (Karlin, 1968), stated as follows: suppose a gambler has m dollars initially. In each of a sequence of rounds, he loses 1 dollar with probability $q \neq \frac{1}{2}$ and wins 1 dollar with probability $1-q$. He stops playing when he has either $m+1$ dollars or has no money left. We have the following result, with a proof available on Page 73 of Karlin (1968).

Lemma 0.1 (Gambler's Ruin Lemma)**.** *In the gambler's ruin problem: (1) the probability that the gambler reaches $m+1$ dollars before reaching 0 dollars is $q_m = \frac{\left(\frac{1-q}{q}\right)^m - 1}{\left(\frac{1-q}{q}\right)^{m+1} - 1}$; (2) the expected number of steps before the gambler stops playing is $\frac{m}{1-2q} - \frac{m+1}{1-2q} \frac{\left(\frac{1-q}{q}\right)^m - 1}{\left(\frac{1-q}{q}\right)^{m+1} - 1}$.*

Observe that the conditional distribution of $T_{\ell,k}$ and the winner of iteration k round ℓ, given the two arms being pulled, is given by the result above for the Gambler's ruin problem. We leverage this in our proof.

1. Proof of Lemma 1

Proof. Suppose we are comparing arm i versus arm j in this iteration with $i > j$ and arm i is the incumbent. Then we know $C(t_{\ell,k}-1, i) = (N-1)(\ell-1)+k-1$ and $C(t_{\ell,k}-1, j) = -\ell+1$. We will keep playing these two arms until $C(t_{\ell,k}+T_{\ell,k}-1, i) = (N-1)(\ell-1)+k$ or $C(t_{\ell,k}+T_{\ell,k}-1, j) = (N-1)(\ell-1)+k$. Further, since the winning probability of arm i over arm j is $p_{i,j}$ over this period, we know the dynamics of this iteration are the same as those of the Gambler's Ruin problem. Denote $E = C(t_{\ell,k}-1, i) - C(t_{\ell,k}-1, j) + 1 = Nl + k - N$. Then the expected length of time we spend in this iteration by Lemma 0.1 is

$$
\begin{aligned}
&\frac{E}{1-2p_{i,j}} - \frac{E+1}{1-2p_{i,j}} \frac{\left(\frac{1-p_{i,j}}{p_{i,j}}\right)^E - 1}{\left(\frac{1-p_{i,j}}{p_{i,j}}\right)^{E+1} - 1} \\
\leq &\frac{E}{1-2p_{i,j}} \leq \frac{E}{2p-1}.
\end{aligned}
$$

The proof of second statement is similar. Using the same notation but now supposing $p_{i,j} \geq p > \frac{1}{2}$, we have that the expected length of time we spend in this iteration is

$$
\begin{aligned}
&\frac{E}{1-2p_{i,j}} - \frac{E+1}{1-2p_{i,j}} \frac{\left(\frac{1-p_{i,j}}{p_{i,j}}\right)^E - 1}{\left(\frac{1-p_{i,j}}{p_{i,j}}\right)^{E+1} - 1} \\
= &\frac{1}{2p_{i,j}-1} - \frac{E+1}{1-2p_{i,j}} \frac{p_{i,j}(1-p_{i,j})^E - (1-p_{i,j})^{E+1}}{(1-p_{i,j})^{n+1} - p_{i,j}^{E+1}} \\
\leq &\frac{1}{2p-1}.
\end{aligned}
$$

□

2. Proof of Lemma 2

In this section, we prove Lemma 2 from the main paper. This section is structured as follows: In section 2.1, we provide two bounds for the incumbent's losing and winning probability; In section 2.2, we consider a version of the problem in which better and worse incumbents have constant (but different) winning probabilities and provide a upper bound for the number of worse incumbents in a round before a better incumbent loses ; In section 2.3, we use the results from the previous subsection to bound the expected number of iterations with a worse incumbent in a single round before a better incumbent loses, starting from within a round; In section 2.4, we prove a similar bound on the expected number of iterations with a worse incumbent in this and future rounds before a better incumbent loses, starting from the beginning of a round; In section 2.5, we complete the proof of Lemma 2.

Throughout this section, we use a one to one correspondence between n and (ℓ, k) defined by $n = (\ell-1)(N-1)+k$, $0 \leq k \leq N-1$ and $\ell = \lceil n/(N-1) \rceil$. We also denote $p^* = \frac{2p-1}{p}$.

2.1. Bounds on Win and Loss Probabilities

We first prove the following two lemmas, which give

- a lower bound for the probability that a worse incumbent loses an iteration;
- an upper bound for the probability that a better incumbent loses an iteration.

Lemma 2.1. *In iteration k of round ℓ conditioned on the identities of the incumbent and the challenger, if the incumbent is worse than the challenger, then the incumbent loses the iteration with conditional probability at least $p^* = \frac{2p-1}{p}$.*

Proof. Let i be the incumbent and j be the challenger, with $i > j$. $C(i, t_{\ell,k}) \geq 0$ and $C(j, t_{\ell,k}) \leq 0$. Let $E = C(i, t_{\ell,k}) + |C(j, t_{\ell,k})| + 1$. The probability that arm i loses this iterations is the same as $1 - q_E$ in the Gambler's Ruin Lemma, Lemma 0.1, with $q = p_{i,j} < 0.5$. This probability is:

$$\begin{aligned}
1 - q_E &= 1 - \frac{\left(\frac{1-p_{j,i}}{p_{i,j}}\right)^E - 1}{\left(\frac{1-p_{i,j}}{p_{i,j}}\right)^{E+1} - 1} \\
&\geq \frac{\left(\frac{1-p_{i,j}}{p_{i,j}}\right)^{E+1} - \left(\frac{1-p_{i,j}}{p_{i,j}}\right)^{E}}{\left(\frac{1-p_{i,j}}{p_{i,j}}\right)^{E+1}} = \frac{1-2p_{i,j}}{1-p_{i,j}} \\
&\geq \frac{2p-1}{p}.
\end{aligned}$$

□

Lemma 2.2. *In iteration k of round ℓ conditioned on the identities of the incumbent and the challenger, if the incumbent is better than the challenger, then the incumbent loses the iteration with conditional probability at most $\left(\frac{1-p}{p}\right)^E$, where $E = N(\ell-1)+k$.*

Proof. This proof is similar to the previous one. Suppose we are pulling arm i and j with $i < j$ and i is the incumbent. Then we know $C(t_{\ell,k}-1, i) = (N-1)(\ell-1)+k-1$ and $C(t_{\ell,k}-1, j) = -\ell+1$. The probability that arm i loses is equal to $1 - q_E$ from the gambler's ruin problem, where $E = (N-1)(\ell-1)+k-1+\ell-1 = N(\ell-1)+k$. We have

$$\begin{aligned}
1 - q_E &= 1 - \frac{\left(\frac{1-p_{i,j}}{p_{i,j}}\right)^E - 1}{\left(\frac{1-p_{i,j}}{p_{i,j}}\right)^{E+1} - 1} \\
&= \frac{\left(\frac{1-p_{i,j}}{p_{i,j}}\right)^E [1 - \frac{1-p}{p}]}{1 - \left(\frac{1-p_{i,j}}{p_{i,j}}\right)^{E+1}} \\
&\leq \left(\frac{1-p_{i,j}}{p_{i,j}}\right)^E \leq \left(\frac{1-p}{p}\right)^E.
\end{aligned}$$

□

2.2. Definition and Upper Bound for $g(b, m)$

In this section, we define a function $g(b, m)$ as follows. First, we define $g(0, m) = 0$ for any m. We define $g(b, m)$ for other integers b, m satisfying $m > 0$ and $0 \leq b \leq m$ recursively, as follows:

$$\begin{aligned}
&g(b, m) \\
=&\frac{b}{m} + \sum_{b'=0}^{b-1} \frac{1}{m} p^* g(b^{'}, m-1) + \sum_{b'=b}^{m-1} \frac{1}{m} g(b, m-1) \\
&+ \sum_{b'=0}^{b-1} \frac{1}{m} (1-p^*) g(b-1, m-1) \\
=&\frac{b}{m} + \sum_{b'=0}^{b-1} \frac{1}{m} p^* g(b^{'}, m-1) + \frac{m-b}{m} g(b, m-1) \\
&+ \frac{b}{m} (1-p^*) g(b-1, m-1) \qquad (1)
\end{aligned}$$

Intuitively, $g(b, m)$ is the expected number of future iterations in which the incumbent is worse than the challenger, starting with m arms that have not dueled yet b of which are better than the incumbent, when we stop counting when we reach the end of the round or when an incumbent loses to a worse challenger, in a simplified problem in which worse incumbents beat better challengers with probability p^*. In our problem, this probability is not p^*, but is bounded below by this quantity, and in the next section we will show that $g(b, m)$ is an upper bound on an analogous quantity in our problem.

We prove the following result about g.

Lemma 2.3. *For $0 \leq b \leq m \leq N-1$, we have*

$$g(b, m) = g(b, b) \leq \frac{\log(b)+1}{p^*}.$$

Proof. Given the boundary conditions $g(0,m)=0$ for all m, we know Equation 1 has a unique solution. In this proof,

- We first assume $g(b,m)=g(b,b)$ for all $b \le m$ and solve for $g(b,m)$;
- Then we show that this $g(b,m)$ is indeed the solution for Equation 1, verifying that $g(b,m)$ is as claimed;
- Finally, we show $g(b,m) \le \frac{\log(b)+1}{p^*}$.

First, we solve for $g(b,m)$ with the assumption that $g(b,m)=g(b,b)$ for $b \le m$. Setting $m=b$ in Equation 1 provides

$$g(b,b) = 1 + \sum_{b'=0}^{b-1} \frac{p^* g(b',b)}{b} + (1-p^*)g(b-1,b-1). \tag{2}$$

Thus, we know

$$\begin{aligned}
&\sum_{b'=1}^{b-1} p^* g(b',b+1)\\
&\quad= \sum_{b'=1}^{b-1} p^* g(b',b)\\
&\quad= b\left[g(b,b) - 1 - (1-p^*)g(b-1,b-1)\right].
\end{aligned}$$

Therefore, Equation 2 becomes

$$\begin{aligned}
&g(b+1,b+1)\\
&=1 + \frac{b}{b+1}[g(b,b) - 1 - (1-p^*)g(b-1,b-1)]\\
&\quad + \frac{p^* g(b,b)}{b+1} + (1-p^* g(b,b).
\end{aligned}$$

Re-organizing the terms, we have

$$\begin{aligned}
&g(b+1,b+1) - g(b,b)\\
&=\frac{1}{b+1} + \frac{b}{b+1}(1-p^*)[g(b,b) - g(b-1,b-1)].
\end{aligned}$$

Denote $F(b) = g(b,b) - g(b-1,b-1)$. We know $F(1) = 1$. Thus, we have

$$\begin{aligned}
F(b) &= \frac{1}{b} + \frac{b-1}{b}(1-p^*)F(b-1)\\
&= \frac{1}{b} + \frac{1-p^*}{b} + \frac{b-2}{b}(1-p^*)^2 F(b-2)\\
&= \frac{1}{b} + \frac{1-p^*}{b} + \cdots \frac{(1-p^*)^{b-1}}{b}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
g(b,b) &= \sum_{k=1}^{b} F(k)\\
&= \sum_{k=1}^{b}\left[\frac{1}{k} + \frac{1-p^*}{k} + \cdots \frac{(1-p^*)^{k-1}}{k}\right].
\end{aligned}$$

Thus, if $g(b,m)=g(b,b)$ for all $b \le m$, we know

$$g(b,m) = \sum_{k=1}^{b}\left[\frac{1}{k} + \frac{1-p^*}{k} + \cdots \frac{(1-p^*)^{k-1}}{k}\right].$$

Now we verify that this is the correct solution. We prove this by induction on b. For $b=1$, Equation 1 becomes

$$g(1,m) = \frac{1}{m} + \frac{m-1}{m} g(1,m-1).$$

Since $g(1,1)=1$, it is easy to check $g(1,2)=g(1,3)=\cdots=g(1,N-1)=1$.

Suppose this $g(b,m)=g(b,b)$ are true for all $b \le m$, $b \le k$. For $b=k+1$, Equation 1 becomes

$$\begin{aligned}
&g(k+1,m)\\
=&\frac{k+1}{m} + \sum_{b'=0}^{k} \frac{p^*}{m} g(b',m-1) + \frac{m-k-1}{m} g(k+1,m-1)\\
&+ \frac{k+1}{m}(1-p^*)g(k,m-1)\\
=&\frac{k+1}{m} + \sum_{b'=0}^{k} \frac{p^*}{m} g(b',b') + \frac{m-k-1}{m} g(k+1,m-1)\\
&+ \frac{k+1}{m}(1-p^*)g(k,k).
\end{aligned}$$

To show $g(k+1,m)$ does not depend on m, we need to prove the following equation is true for $m=k+2, k+3, \cdots, N-1$.

$$\begin{aligned}
&\frac{k+1}{m} + \sum_{b'=0}^{k} \frac{p^*}{m} g(b',b') + \frac{k+1}{m}(1-p^*)g(k,k)\\
&=\frac{k+1}{m} g(k+1,m-1)\\
\iff& k+1 + \sum_{b'=0}^{k} p^* g(b',b') + (k+1)(1-p^*)g(k,k)\\
&=(k+1)g(k+1,m-1)
\end{aligned} \tag{3}$$

We first check Equation 3 when $m=k+2$. Starting from the left hand side, we have

$$\begin{aligned}&k+1+\sum_{b^{'}=0}^{k} g(b^{'},b^{'})+(k+1)(1-p^*)g(k,k)\\ =&k+1+(k+1)[g(k+1,k+1)-1-(1-p^*)g(k,k)] \qquad (4)\\ &+(k+1)(1-p^*)g(k,k)\\ =&(k+1)g(k+1,k+1),\end{aligned}$$

which equals to the right hand side. Equation (4) follows from Equation (2) (Equation (2) holds because $g(b,m)=g(b,b)$ for all $b \leq k$).

Again, by induction, we know (3) is true for all $m = k+2, \cdots, N-1$ and thus we concludes our induction.

We have shown that $g(b,m)=g(b,b)$ for all $b \leq m$.

Finally, we prove $g(b,b)=g(b,m) \leq \frac{\log(b)+1}{p^*}$. This is because

$$\begin{aligned}g(b,m)=&g(b,b)\\ =&\sum_{k=1}^{b}\left[\frac{1}{k}+\frac{1-p^*}{k}+\cdots\frac{(1-p^*)^{k-1}}{k}\right]\\ \leq&\sum_{k=1}^{b}\left[\frac{1}{k}+\frac{1-p^*}{k}+\cdots\frac{(1-p^*)^{b-1}}{k}\right]\\ =&\sum_{k=1}^{b}\frac{1}{k}\left[1+(1-p^*)+\cdots+(1-p^*)^{b-1}\right]\\ \leq&\frac{\log(b)+1}{p^*},\end{aligned}$$

which concludes our proof. □

2.3. Bound on the Number of Iterations in One Round with a Worse Incumbent, Starting from Within the Round

Let $B(n)$ denote an indicator function that equals 1 if we have a better incumbent at the n^{th} iteration. The definition of $B(n)$ is very similar to $B(\ell,k)$ except $B(\ell,k)$ tracks both round and iteration number. Similarly, we use $\bar{B}(n)=1-B(n)$ to denote an indicator function that equals 1 if we have a worse incumbent at the n^{th} iteration.

Let $h(i,n,\mathcal{A})$ be the expected number of iterations with an incumbent that is worse than the challenger, between iteration n and the first time that a better incumbent loses to a challenger or the round ends, given that the incumbent arm at iteration n is i and $\mathcal{A}$ is the set of arms that have not yet previously dueled in the round. Formally, we define this quantity as:

$$h(i,n,\mathcal{A})=\mathbb{E}\left[\sum_{n'=n}^{\sigma-1}B(n')|\mathcal{A},i_n=i\right],$$

where

- Conditioning on $\mathcal{A}$ is understood to mean that we are conditoning on $C(n-1,j)=-\ell+1\ \forall\ j\notin \mathcal{A}\cup\{i_n\}$, and $C(n-1,j)=-\ell\ \forall\ j\in\mathcal{A}$, where $\ell=\lceil n/(N-1)\rceil$ is the round in which iteration n resides. In other words, it is understood to mean that $\mathcal{A}$ contains the set of arms that have not yet dueled in this round.
- $\sigma=\min\{n'>n: J(n')=1, n'=N\lceil n/(N-1)\rceil\}$ where $J(n)$ is an indicator that equals 1 when a better incumbent loses at iteration n, i.e., σ is the first time that either a better incumbent loses or the round ends.

Lemma 2.4. *For any i, ℓ, k and $\mathcal{A}$, we have*

$$h(i,n,\mathcal{A})\leq g(b,m)\leq\frac{\log(N)+1}{p^*},$$

where $m=N-k$ and $b=|\{u\in\mathcal{A}:u<i\}|$.

Proof. Denote $q_{i,j}(n)$ as the probability that incumbent arm i will beat challenger j at time n. We first write a recursive expression for $h(i,n,\mathcal{A})$ that applies when n is not divisible by N:

$$\begin{aligned}h(i,n,\mathcal{A})=&\sum_{\{j\in\mathcal{A}:i>j\}}\Bigg[1+\frac{q_{i,j}(n)}{N-k}h(i,n+1,\mathcal{A}\cup\{j\})\\ &+\frac{1-q_{i,j}(n)}{N-k}h(j,n+1,\mathcal{A}\cup\{i\})\Bigg]\\ &+\sum_{\{j\in\mathcal{A}:i<j\}}\frac{q_{i,j}(n)}{N-k}h(i,n+1,\mathcal{A}\cup j). \quad (5)\end{aligned}$$

When n is divisible by $N-1$, the only allowed value of $\mathcal{A}$ is $\emptyset$ and $h(i,n,\emptyset)=0$.

We then prove the desired result via induction on the number of iterations in the round, i.e., on $n \pmod{N-1}$. When $n \pmod{N-1}=0$, we have $h(i,n,\emptyset)=0$, $b=0$, and $g(b,m)=0$. Thus the result holds in this case.

Then suppose the result holds for all n with a particular value of $n \pmod{N-1}$ and we show it holds for $n-1$.

Applying the induction hypothesis to the right-hand side of (5), we have

$$\begin{aligned} h(i,n,\mathcal{A}) \leq & \sum_{\{j\in\mathcal{A}:i>j\}} \Big[1 + \frac{q_{i,j}(n)}{m} g(b_{i,j}, m-1) \\ & + \frac{1-q_{i,j}(n)}{m} g(b_{j,j}, m-1)\Big] \\ & + \sum_{\{j\in\mathcal{A}:i<j\}} \frac{q_{i,j}(n)}{m} g(b_{i,j}, m-1), \qquad (6) \end{aligned}$$

where $b_{u,j} = \#\{u^{'} \in \mathcal{A} \setminus \{j\} : u^{'} < u\}$.

Consider the summand in the first sum in (6), dropping the constants 1 and $\frac{1}{m}$,

$$q_{i,j}(n) g(b_{i,j}, m-1) + (1-q_{i,j}(n)) g(b_{j,j}, m-1). \quad (7)$$

This is increasing in $q_{i,j}(n)$ when $i > j$ since $b_{i,j} > b_{j,j}$, and since $g(b,m)$ is increasing in b. Since i is an incumbent that is worse than the challenger when $i > j$, Lemma 2.1 shows that $q_{i,j}(n) \leq 1 - p^* = 1 - \frac{2p-1}{p}$ in this situation. Thus, this summand is bounded above by $(1-p^*)g(b_{i,j}, m-1) + p^* g(b_{j,j}, m-1)$.

Substituting this into (6), along with the inequality $q_{i,j}(n) \leq 1$ in the last term, we have

$$\begin{aligned} & h(i,n,\mathcal{A}) \\ \leq & \sum_{\{j\in\mathcal{A}:i>j\}} \Big[1 + \frac{1-p^*}{m} g(b_{i,j}, m-1) + \frac{p^*}{m} g(b_{j,j}, m-1)\Big] \\ & + \sum_{\{j\in\mathcal{A}:i<j\}} \frac{1}{m} g(b_{i,j}, m-1) \\ = & \frac{b}{m} + \frac{b}{m}(1-p^*) g(b-1, m-1) + \sum_{b'=0}^{b-1} \frac{p^*}{m} g(b^{'}, m-1) \\ & + \frac{m-b}{m} g(b, m-1) \\ = & g(b,m) \end{aligned}$$

In the second to last line we have used that $\{b_{i,j} : j \in \mathcal{A}, i > j\} = \{0, \ldots, b-1\}$ and $b_{i,j} = b-1$ when $i > j$; $b_{i,j} = b$ when $i < j$; and that the cardinality of $\{j \in \mathcal{A} : i > j\}$ and $\{j \in \mathcal{A} : i < j\}$ are b and $m - b$ respectively. In the last line we have used the recursive definition of $g(b,m)$ in terms of $g(\cdot, m-1)$.

This shows the first inequality in the statement of the lemma. The second inequality follows directly from Lemma 2.3. □

2.4. Bound on the Number of Iterations with a Worse Incumbent, Starting from a Round Beginning

Denote $f(i,\ell)$ to be the expected number of iterations with a worse incumbent in this and future rounds, stopping as soon as a better incumbent loses, giving that we have arm i as the incumbent at the start of round ℓ.

Lemma 2.5. *For any i and ℓ, we have*

$$f(i,\ell) \leq \frac{\log(N)+1}{(p^*)^2}.$$

Proof. Let $U(i,\ell)$ denote the expected number of iterations in round ℓ with a worse incumbent before a better incumbent loses. We use $V(\ell)$ to denote an indicator which equals to 1 if a better incumbent does not lose in the round ℓ. Then for $i > 1$,

$$f(i,\ell) = U(i,\ell) + \mathbb{E}[f(Z(\ell), \ell+1) V(\ell) | Z(\ell-1) = i].$$

The first term is bounded by Lemma 2.4 by

$$U(i,\ell) \leq \frac{\log(N)+1}{p^*},$$

for all i and ℓ.

For the second term, since $f(Z(\ell), \ell+1) = 0$ when $Z(\ell) = 1$, we know the second term is bounded by

$$\begin{aligned} & \mathbb{E}[f(Z(\ell), \ell+1) V(\ell) | Z(\ell-1) = i] \\ \leq & \mathbb{E}[f(Z(\ell), \ell+1) | Z(\ell) \neq 1, V(\ell) = 1, Z(\ell-1) = i] \\ & \times P(Z(\ell) \neq 1, V(\ell) | Z(\ell-1) = i). \end{aligned}$$

Let $s_j = P(Z(\ell) = j | Z(\ell) \neq 1, V(\ell), Z(\ell-1) = i)$ to be the probability distribution over the integers from 2 through N. Then we know

$$\begin{aligned} & \mathbb{E}[f(Z(\ell), \ell+1) | Z(\ell) \neq 1, V(\ell) = 1, Z(\ell-1) = i] \\ = & \sum_{j=2}^{N} s_j f(j, \ell+1) \\ \leq & \max_{j=2,,N} f(j, \ell+1). \end{aligned}$$

Further, since if arm 1 wins its first duel as a challenger (which happens with probability at least p^*), then either $Z(\ell) = 1$ (it wins all subsequent duel in the round) or $V(\ell) = 0$ (it loses a subsequent duel), we have $P(Z(\ell) \neq 1, V(\ell) | Z(\ell-1) = i) \leq 1 - p*$.

Thus, we know

$$f(i,\ell) \leq \frac{\log(N)+1}{p^*} + (1-p^*) \max_{j=2,\cdots,N} f(j, \ell+1).$$

Let $f(\ell) = \max_{j=2,\cdots,N} f(j,\ell)$. Then,

$$f(\ell) \leq \frac{\log(N)+1}{p^*} + (1-p^*) f(\ell+1).$$

Thus,

$$\begin{aligned} f(1) \le& \frac{\log(N)+1}{p^*} + (1-p^*)f(2) \\ \le& \frac{\log(N)+1}{p^*}(1+(1-p^*)+(1-p^*)^2+\cdots) \\ =& \frac{\log(N)+1}{(p^*)^2}. \end{aligned}$$

□

2.5. Completing the Proof of Lemma 3

With the lemmas in the preceding subsections established, we now complete the proof of Lemma 2.

Proof. Let $\tau_0 = 0$ and $\tau_k = \{n > \tau_{k-1} : J(n) = 1\}$. The expected number of iterations with a worse incumbent is

$$\begin{aligned} &\mathbb{E}\left[\sum_{n=0}^{\infty} \bar{B}(n)\right] \\ =&\mathbb{E}\sum_{k=0}^{\infty} 1\{\tau_k < \infty\} \sum_{n=\tau_k}^{\infty} 1\{n < \tau_{k+1}\}\bar{B}(n) \\ =&\sum_{k=0}^{\infty} P(\tau_k < \infty)\mathbb{E}\left[\sum_{n=\tau_k}^{\infty} 1\{n < \tau_{k+1}\}\bar{B}(n)|\tau_k < \infty\right] \end{aligned}$$

where we have used Tonelli's theorem to exchange the expectation of an infinite sum of non-negative terms with an infinite sum of expectations of the same terms.

Conditioning on the history available at time τ_k, we have that the inner expectation can be written as,

$$\begin{aligned} &\mathbb{E}\left[\sum_{n=\tau_k}^{\infty} 1\{n < \tau_{k+1}\}\bar{B}(n)|\tau_k < \infty\right] \\ =&\mathbb{E}\left[\mathbb{E}\left[\sum_{n=\tau_k}^{\infty} 1\{n < \tau_{k+1}\}\bar{B}(n)|H_{\tau_k}, \tau_k < \infty\right]|\tau_k < \infty\right], \end{aligned}$$

where H_n is the sigma algebra generated by $(C(i,s) : s < t_{\ell,k'}, i = 1,\dots,N)$, where $\ell = n \pmod{N-1}$, $k' = \lceil n/(N-1)\rceil$, and H_{τ_k} is the filtration $(H_n : n)$ stopped at τ_k.

We further break this inner term $\mathbb{E}\left[\sum_{n=\tau_k}^{\infty} 1\{n < \tau_{k+1}\}\bar{B}(n)|H_{\tau_k}, \tau_k < \infty\right]$ into two parts: the part that occurs during the round in which τ_k resides, and the part that occurs in future rounds. Let $\ell_k = \lceil \tau_k/(N-1)\rceil$. Then,

$$\begin{aligned} &\mathbb{E}\left[\sum_{n=\tau_k}^{\infty} 1\{n < \tau_{k+1}\}\bar{B}(n)|H_{\tau_k}, \tau_k < \infty\right] \\ =&\mathbb{E}\left[\sum_{n=\tau_k}^{\ell_k N} 1\{n < \tau_{k+1}\}\bar{B}(n)|H_{\tau_k}, \tau_k < \infty\right] \\ +&\mathbb{E}\left[\sum_{n=\ell_k N+1}^{\infty} 1\{n < \tau_{k+1}\}\bar{B}(n)|H_{\tau_k}, \tau_k < \infty\right] \\ \le&\frac{\log(N)+1}{p^*} + \frac{\log(N)+1}{(p^*)^2} \\ \le&\frac{2(\log(N)+1)}{(p^*)^2} \end{aligned}$$

where the second to last inequality relies on Lemma 2.4 to show $\mathbb{E}\left[\sum_{n=\tau_k}^{\ell_k N} 1\{n < \tau_{k+1}\}\bar{B}(n)|H_{\tau_k}, \tau_k < \infty\right]$ is bounded above by $\frac{\log(N)+1}{p^*}$ and Lemma 2.5 to show $\mathbb{E}\left[\sum_{n=\ell_k N+1}^{\infty} 1\{n < \tau_{k+1}\}\bar{B}(n)|H_{\tau_k}, \tau_k < \infty\right]$ is bounded above by $\frac{\log(N)+1}{(p^*)^2}$.

Thus,

$$\mathbb{E}\left[\sum_{n=0}^{\infty} \bar{B}(n)\right] \le \frac{2(\log(N)+1)}{(p^*)^2}\sum_{k=0}^{\infty} P(\tau_k < \infty).$$

Now we bound $P(\tau_k < \infty)$ for a fixed k. Based on Lemma 2.2, we know $J(n)$ is a Bernoulli random variable with success rate less than $\left(\frac{1-p}{p}\right)^n$ (this is because of Lemma 2.2 and $n = (N-1)(\ell-1)+k < E$), independent across n. Let Q_n denote a Bernoulli random variable with success rate $\left(\frac{1-p}{p}\right)^n$. Then we know:

$$\begin{aligned} P(\tau_k < \infty) \le& P\left(\sum_{i=1}^{\infty} J(i) \ge k\right) \\ \le& P\left(\sum_{i=1}^{\infty} Q_i \ge k\right). \end{aligned}$$

Let $W_m = \sum_{i=1}^m Q_i$, which follows a Poisson Bernoulli distribution, and let $W = \lim_{m\to\infty} W_m$. W follows a Poisson distribution with parameter $\sum_{i=1}^{\infty}\left(\frac{1-p}{p}\right)^i = \frac{1-p}{2p-1}$ (Theorem 4, Wang (1993)). Thus,

$$\begin{aligned} \mathbb{E}\left[\sum_{n=0}^{\infty} \bar{B}(n)\right] \le& \frac{2(\log(N)+1)}{(p^*)^2}\sum_{k=0}^{\infty} P(W \ge k) \\ =& \frac{2p^2(1-p)}{(2p-1)^3}(\log(N)+1) \\ \le& \frac{2p^2}{(2p-1)^3}(\log(N)+1) \end{aligned}$$

□

3. Proof of Lemma 3

Proof. It is easy to see that at the last iteration which has a worse incumbent, the better arm is always arm 1. Thus, we only consider $C(t,1)$ in this proof. At the end of the ℓ^{th} round, if $C(t_{\ell+1}-1,1)<0$, we know $C(t_{\ell+1}-1,1)=-\ell$.

Let us consider a simple random walk W(t) such that $W(t+1)=W(t)+1$ with probability $p>\frac{1}{2}$ and $W(t+1)=W(t)-1$ with probability $1-p$ for $t\geq 1$. If we denote $p^*_\ell = P(\exists t_*, W(t_*)=-\ell)$ for $\ell>0$, then it is easy to calculate that $p^*_\ell=\left(\frac{1-p}{p}\right)^\ell$.

Now let us consider $C(t,1)$. If we pull arm 1 with some other arm i at time t, then $C(t,1)=C(t-1,1)+1$ happens with probability $p_{1,i}>p$ and $C(t,1)=C(t-1,1)-1$ with probability $1-p_{1,i}<1-p$. If we do not pull arm 1 at time t, then $C(t,1)=C(t-1,1)$ with probability 1.

Define $\tau_1=1$ and $\tau_k=\min_t\{t>\tau_{k-1}, C(t,1)\neq C(\tau_{k-1},1)\}$, for $k=1,2,\cdots,$. Because τ_k is a non-decreasing right continuous stopping time, we know it is a valid random change of time (Barndorff-Nielsen & Shiryaev, 2015). Define $R(k)$ a new stochastic process where $R(k)=C(\tau_k,1)$. Then we know at every time k, $R(k)=R(k-1)+1$ with probability greater or equal to p and $R(k)=R(k-1)-1$ with probability less than 1-p. Define $p_\ell=P(\exists t_*, R(t_*)=-\ell)$, then it is easy to prove $p_\ell\leq p^*_\ell=\left(\frac{1-p}{p}\right)^\ell$ using first step analysis and induction (we leave the proof as an exercise for the reader), which means $P(\exists t_*, C(t_*,1)=-\ell)\leq\left(\frac{1-p}{p}\right)^\ell$. □

4. Proof of Lemma 4

Proof. To show the first claimed equation, we have:

$$\begin{aligned}&\mathbb{E}[B(\ell,k)T_{\ell,k}\bar{D}(\ell)]\\ =&\mathbb{E}[B(\ell,k)T_{\ell,k}|\bar{D}(\ell)=1]P(\bar{D}(\ell)=1).\end{aligned}\tag{8}$$

The first term $\mathbb{E}[B(\ell,k)T_{\ell,k}|\bar{D}(\ell)=1]$ can be bounded by writing it as $\mathbb{E}[B(\ell,k)T_{\ell,k}|\bar{D}(\ell)=1]=\mathbb{E}[\mathbb{E}[B(\ell,k)T_{\ell,k}|\bar{D}(\ell)=1,A(\ell,k)]|\bar{D}(\ell)=1]$, where $A(\ell,k)$ denotes the pair of arms being pulled in iteration k round ℓ.

We focus on the inner term $\mathbb{E}[B(\ell,k)T_{\ell,k}|\bar{D}(\ell)=1,A(\ell,k)]$. $B(\ell,k)$ is observable given $A(\ell,k)$. If $B(\ell,k)=0$ then this inner term is 0. If $B(\ell,k)=1$ then this inner term is $\mathbb{E}[T_{\ell,k}|A(\ell,k)]$ (where we note that $T_{\ell,k}$ is conditionally independent of $\bar{D}(\ell)$ given $A(\ell,k)$) and is bounded above by $1/(2p-1)$ by Lemma 1. In both cases, the inner term is bounded above by $1/(2p-1)$, and we have that $\mathbb{E}[B(\ell,k)T_{\ell,k}|\bar{D}(\ell)=1]\leq 1/(2p-1)$.

Thus, we have that (8) is bounded above by

$$\frac{1}{2p-1}P(\bar{D}(\ell)=1)\leq\frac{1}{2p-1}\left(\frac{1-p}{p}\right)^{\ell-1},$$

where the final inequality follows from Lemma 3 and the fact that $\bar{D}(\ell)=1$ implies $L\geq\ell-1$.

To show the second claimed equation, we use the same proof technique used for the first and get:

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

Now we just need to compute $P(V(\ell,k)=1)$. Given $C(t_\ell-1,1)=(N-1)(\ell-1)$ at the beginning of round ℓ, it loses only if there exists a $t_0\geq t_\ell$ and $C(1,t_0)=-\ell$. Using the results from Lemma 3, we know $P(V(\ell,k)=1)\leq\left(\frac{1-p}{p}\right)^\ell$. This completes the proof of the second claimed equation. □

5. Proof of Lemma 5

Proof. For the first inequality, we know

$$\begin{aligned}&\mathbb{E}\left[\sum_{k=1}^{N-1}\bar{B}(\ell,k)T_{\ell,k}\bar{D}(\ell)\right]\\ &=\sum_{k=1}^{N-1}\mathbb{E}\left[\mathbb{E}[\bar{B}(\ell,k)T_{\ell,k}|D(\ell)=0]\bar{D}(\ell)\right].\end{aligned}\tag{9}$$

Moreover,

$$\begin{aligned}&\mathbb{E}[\bar{B}(\ell,k)T_{\ell,k}|D(\ell)=0]\\ =&\mathbb{E}[T_{\ell,k}|B(\ell,k)=0,D(\ell)=0]P(B(\ell,k)=0|D(\ell)=0)\\ \leq&\frac{N\ell}{2p-1}P(B(\ell,k)=0|D(\ell)=0),\end{aligned}$$

where the last equation follows from applying Lemma 1 and iterated conditional expectation. Thus, we know

$$\begin{aligned}(9)&=\sum_{k=1}^{N-1}\frac{N\ell}{2p-1}P(B(\ell,k)=0|D(\ell)=0)\mathbb{E}[\bar{D}(\ell)]\\ &\leq\sum_{k=1}^{N-1}\frac{N\ell}{2p-1}P(B(\ell,k)=0|D(\ell)=0)\left(\frac{1-p}{p}\right)^{\ell-1} \qquad (10)\\ &\leq\left(\frac{1-p}{p}\right)^{\ell-1}\frac{2N\ell p^2}{(2p-1)^4}(\log(N)+1),\end{aligned}$$

where equation (10) is because Lemma 2. □

The proof of the second inequality follows very similarly, and is omitted. □

6. Proof of Theorem 2

In this section, we prove the cumulative expected weak regret of WS-W is bounded by $O(N^2)$ in the Condorcet winner setting. First, we want to give an example to illustrate why our algorithm will not have $O(N\log(N))$ regret under the Condorcet winner setting.

In the Condorcet winner setting, Lemma 2 is no longer true. Here is a counter example to illustrate why Lemma 2 does not hold true anymore. Suppose we have $N = 3k + 1$ arms in total, which includes a Condorcet winner arm and three types of other arms: k type-A arms, k type-B arms and k type-C arms. Among these arms, we assume the user prefers type-A arms than type-B arms, type-B arms than type-C arms and type-C arms than type-A arms. Among each type of arms, there is a total order. In this setting, the expected number of iterations with a worse incumbent is $O(N)$ instead of $O(\log(N))$, which means Lemma 2 is no longer true.

Now we start our proof for Theorem 2.

Proof. In the Condorcent winner setting, Lemmas 3 and 4 hold, but as explained earlier, Lemma 2 does not. Because the proof of Lemma 5 utilizes Lemma 2, Lemma 5 also no longer holds.

On the other hand, since we can have at most $N - 1$ iterations in a round, we know the following statement is true: the conditional expected number of iterations with a worse incumbent is bounded by N in each round. Thus, we know Lemma 5 now becomes:

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

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

Thus, following the same reasoning as in the proof of Theorem 1, we know the expected weak regret in the Condorcet winner setting is bounded by

$$\frac{NR}{(2p-1)^2} + \frac{pN^2}{(2p-1)^3},$$

which concludes our proof.

7. Preference Matrices

In the sushi experiment, the user's preference matrix is given by Figure 1.

In the MSLR experiment, the ranker's preference matrix is given by:

$$\begin{bmatrix} 0.5 & 0.535 & 0.613 & 0.757 & 0.765 \\ 0.465 & 0.5 & 0.580 & 0.727 & 0.738 \\ 0.387 & 0.420 & 0.5 & 0.659 & 0.669 \\ 0.243 & 0.276 & 0.341 & 0.5 & 0.510 \\ 0.235 & 0.262 & 0.331 & 0.490 & 0.5 \end{bmatrix}$$

8. Condorcet Winner Experiment

In the main paper, we considered numerical examples in which the arms have a total order. This is common in the dueling bandits literature, where even work that considers more general settings theoretically test their methods on problems that satisfy the total order assumption (Komiyama et al., 2016; Urvoy et al., 2013).

In this section, we consider an additional example that has a Condorcet winner but does not have a total order among arms. The example has a cyclic struture, and is similar to the cyclic example in Komiyama et al. (2015).

The preference matrix is:

$$\begin{bmatrix} 0.5 & 0.6 & 0.6 & 0.6 \\ 0.4 & 0.5 & 0.6 & 0.4 \\ 0.4 & 0.4 & 0.5 & 0.6 \\ 0.4 & 0.6 & 0.4 & 0.5 \end{bmatrix}$$

In the above example, arm 1 is the Condorcet winner. Arm 2 beats arm 3, arm 3 beats arm 4 and arm 4 beats arm 2.

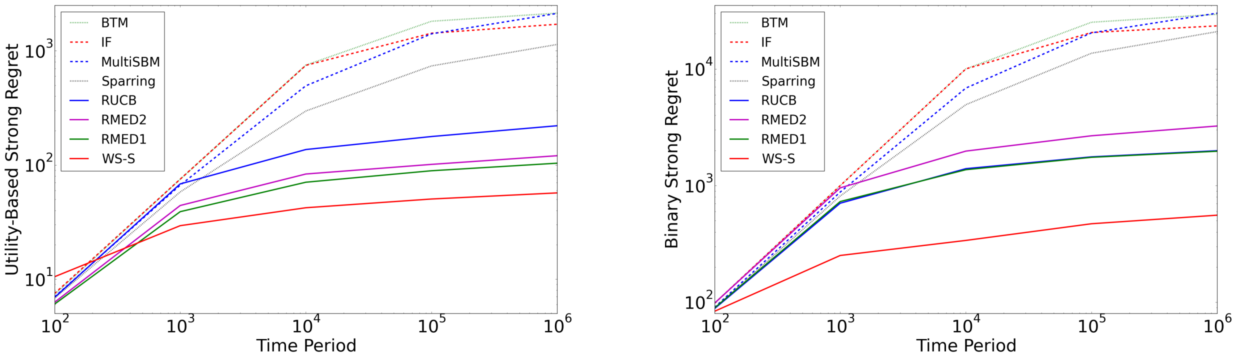
Again, we consider both binary strong regret and the utility-based strong regret. The utility-based strong regret is defined the same as the other two experiments. The result is summarized in Figure 2. WS-S outperforms all benchmarks considered in all time periods on binary regret, and outperforms them all in all time periods except $T = 10^2$ on utility-based regret.

9. Sensitivity Analysis

In this section, we conduct a sensitivity analysis of β in WS-S using the MSLR dataset. In this analysis,

$$\begin{bmatrix}
0.5 & 0.512 & 0.622 & 0.655 & 0.698 & 0.726 & 0.711 & 0.708 & 0.749 & 0.8 & 0.741 & 0.783 & 0.847 & 0.817 & 0.854 & 0.868 \\
0.488 & 0.5 & 0.602 & 0.683 & 0.652 & 0.776 & 0.663 & 0.683 & 0.738 & 0.709 & 0.786 & 0.802 & 0.83 & 0.85 & 0.871 & 0.873 \\
0.378 & 0.398 & 0.5 & 0.528 & 0.554 & 0.533 & 0.534 & 0.591 & 0.573 & 0.593 & 0.661 & 0.705 & 0.734 & 0.672 & 0.787 & 0.822 \\
0.345 & 0.317 & 0.472 & 0.5 & 0.553 & 0.619 & 0.566 & 0.641 & 0.675 & 0.687 & 0.665 & 0.696 & 0.803 & 0.823 & 0.796 & 0.844 \\
0.302 & 0.348 & 0.446 & 0.447 & 0.5 & 0.513 & 0.524 & 0.518 & 0.608 & 0.538 & 0.643 & 0.61 & 0.695 & 0.672 & 0.681 & 0.775 \\
0.274 & 0.224 & 0.467 & 0.381 & 0.487 & 0.5 & 0.513 & 0.559 & 0.575 & 0.621 & 0.591 & 0.701 & 0.702 & 0.787 & 0.829 & 0.811 \\
0.289 & 0.337 & 0.466 & 0.434 & 0.476 & 0.487 & 0.5 & 0.559 & 0.553 & 0.613 & 0.564 & 0.607 & 0.703 & 0.735 & 0.736 & 0.801 \\
0.292 & 0.317 & 0.409 & 0.359 & 0.482 & 0.441 & 0.441 & 0.5 & 0.556 & 0.527 & 0.562 & 0.58 & 0.668 & 0.805 & 0.777 & 0.767 \\
0.251 & 0.262 & 0.427 & 0.325 & 0.392 & 0.425 & 0.447 & 0.444 & 0.5 & 0.512 & 0.548 & 0.542 & 0.612 & 0.786 & 0.71 & 0.685 \\
0.2 & 0.291 & 0.407 & 0.313 & 0.462 & 0.379 & 0.387 & 0.473 & 0.488 & 0.5 & 0.543 & 0.579 & 0.613 & 0.718 & 0.685 & 0.747 \\
0.259 & 0.214 & 0.339 & 0.335 & 0.357 & 0.409 & 0.436 & 0.438 & 0.452 & 0.457 & 0.5 & 0.564 & 0.625 & 0.618 & 0.702 & 0.684 \\
0.217 & 0.198 & 0.295 & 0.304 & 0.39 & 0.299 & 0.393 & 0.42 & 0.458 & 0.421 & 0.436 & 0.5 & 0.542 & 0.644 & 0.7 & 0.733 \\
0.153 & 0.17 & 0.266 & 0.197 & 0.305 & 0.298 & 0.297 & 0.332 & 0.388 & 0.387 & 0.375 & 0.458 & 0.5 & 0.577 & 0.607 & 0.596 \\
0.183 & 0.15 & 0.328 & 0.177 & 0.328 & 0.213 & 0.265 & 0.195 & 0.214 & 0.282 & 0.382 & 0.356 & 0.423 & 0.5 & 0.578 & 0.637 \\
0.146 & 0.129 & 0.213 & 0.204 & 0.319 & 0.171 & 0.264 & 0.223 & 0.29 & 0.315 & 0.298 & 0.3 & 0.393 & 0.422 & 0.5 & 0.586 \\
0.132 & 0.127 & 0.178 & 0.156 & 0.225 & 0.189 & 0.199 & 0.233 & 0.315 & 0.253 & 0.316 & 0.267 & 0.404 & 0.363 & 0.414 & 0.5
\end{bmatrix}$$

Figure 1: User's preference matrix for the Sushi experiment



(a) Cyclic dataset with utility-based strong regret

(b) Cyclic dataset with binary strong regret

Figure 2: Comparison of the strong regret between WS-S and 7 benchmarks on the cyclic dataset. WS-S outperforms all benchmarks in all settings studied.

we choose $\beta = 1.01, 1.05, 1.1, 1.2, 1.5$ respectively and compare them with RMED and RUCB. The result is summarized in Figure 3.

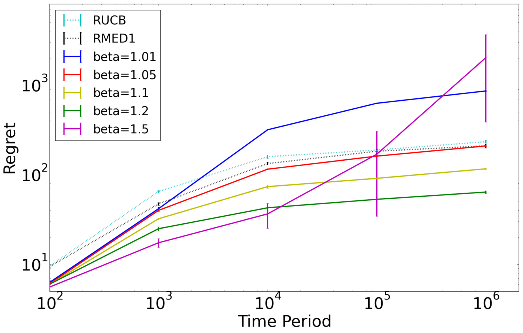


Figure 3: Sensitivity Analysis

Based on Figure 3, WS-S with $\beta = 1.05, 1.1, 1.2$ outperforms RMED and RUCB. When $\beta = 1.01$, we spend too much time on the exploration period and do not exploit enough. Similarly, WS-S with $\beta = 1.5$ over exploits and does not explore enough. In both cases, WS-S underperforms RMED and RUCB. However, as long as β is within a reasonable range, WS-S can outperform existing state-of-art algorithms.

References

Barndorff-Nielsen, Ole E and Shiryaev, Albert. *Change of time and change of measure*, volume 21. World Scientific Publishing Co Inc, 2015.

Karlin, Samuel. *A First Course In Stochastic Processes*. Academic Press, 1968.

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.

Urvoy, Tanguy, Clerot, Fabrice, Féraud, Raphael, and Naamane, Sami. Generic exploration and k-armed voting bandits. In *ICML (2)*, pp. 91–99, 2013.

Wang, Y.H. On the number of success in independent trials. *Statistica Sinica*, 1993.