Beat the Mean Bandit

Yisong Yue YISONGYUE@CMU.EDU
H. John Heinz III College, Carnegie Mellon University, Pittsburgh, PA, USA

Thorsten Joachims TJ@CS.CORNELL.EDU
Department of Computer Science, Cornell University, Ithaca, NY, USA

Abstract

The Dueling Bandits Problem is an online learning framework in which actions are restricted to noisy comparisons between pairs of strategies (also called bandits). It models settings where absolute rewards are difficult to elicit but pairwise preferences are readily available. In this paper, we extend the Dueling Bandits Problem to a relaxed setting where preference magnitudes can violate transitivity. We present the first algorithm for this more general Dueling Bandits Problem and provide theoretical guarantees in both the online and the PAC settings. We also show that the new algorithm has stronger guarantees than existing results even in the original Dueling Bandits Problem, which we validate empirically.

1. Introduction

Online learning approaches have become increasingly popular for modeling recommendation systems that learn from user feedback. Unfortunately, conventional online learning methods assume that absolute rewards (e.g. rate A from 1 to 5) are observable and reliable, which is not the case in search engines and other systems that have access only to implicit feedback (e.g. clicks) (Radlinski et al., 2008). However, for search engines there exist reliable methods for inferring preference feedback (e.g. is A better than B) from clicks (Radlinski et al., 2008). This motivates the K-armed Dueling Bandits Problem (Yue et al., 2009), which formalizes the problem of online learning with preference feedback instead of absolute rewards.

One major limitation of existing algorithms for the Dueling Bandits Problem is the assumption that user preferences satisfy strong transitivity. For example, for strategies A, B and C, if users prefer A to B by 55%, and B to C by 60%, then strong transitivity requires that users prefer A to C at least 60%. Such requirements are often violated in practice (see Section 3.1).

In this paper, we extend the K-armed Dueling Bandits Problem to a relaxed setting where stochastic preferences can violate strong transitivity. We present a new algorithm, called "Beat-the-Mean", with theoretical guarantees that are not only stronger than previous results for the original setting, but also degrade gracefully with the degree of transitivity violation. We empirically validate our findings and observe that the new algorithm is indeed more robust, and that it has orders-of-magnitude lower variability. Finally, we show that the new algorithm also has PAC-style guarantees for the Dueling Bandits Problem.

2. Related Work

Conventional multi-armed bandit problems have been well studied in both the online (Lai & Robbins, 1985; Auer et al., 2002) and PAC (Mannor & Tsitsiklis, 2004; Even-Dar et al., 2006; Kalyanakrishnan & Stone, 2010) settings. These settings differ from ours primarily in that feedback is measured on an absolute scale.

Methods that learn using noisy pairwise comparisons include active learning approaches (Radlinski & Joachims, 2007) and algorithms for finding the maximum element (Feige et al., 1994). The latter setting is similar to our PAC setting, but requires a common stochastic model for all comparisons. In contrast, our analysis explicitly accounts for not only that different pairs of items yield different stochastic preferences, but also that these preferences might not be internally consistent (i.e. violate strong transitivity).

Yue & Joachims (2009) considered a continuous version of the Dueling Bandits Problem, where bandits

Appearing in *Proceedings of the 28th International Conference on Machine Learning*, Bellevue, WA, USA, 2011.

Table 1. $\Pr(Row > Col) - 1/2$ as estimated from interleaving experiments with six retrieval functions on ArXiv.org.

	A	B	C	D	E	F
A	0	0.05	0.05	0.04	0.11	011
B	-0.05	0	0.05	0.06	0.08	0.10
C	-0.05	-0.05	0	0.04	0.01	0.06
D	-0.04	-0.04	-0.04	0	0.04	0.00
E	-0.11	-0.08	-0.01	-0.04	0	0.01
F	-0.11	-0.10	-0.06	-0.00	-0.01	0

are represented as high dimensional points, and derivatives are estimated by comparing two bandits. More recently, Agarwal et al. (2010) proposed a near-optimal multi-point algorithm for the conventional continuous bandit setting, which may be adaptable to the Dueling Bandits Problem.

Our proposed algorithm is structurally similar to the Successive Elimination algorithm proposed by Even-Dar et al. (2006) for the conventional PAC bandit setting. Our theoretical analysis differs significantly due to having to deal with pairwise comparisons.

3. The Learning Problem

The K-armed Dueling Bandits Problem (Yue et al., 2009) is an iterative learning problem on a set of bandits $\mathcal{B} = \{b_1, \dots, b_K\}$ (also called arms or strategies). Each iteration comprises a noisy comparison (duel) between two bandits (possibly the same bandit with itself). We assume the comparison outcomes to have independent and time-stationary distributions.

We write the comparison probabilities as $P(b > b') = \epsilon(b, b') + 1/2$, where $\epsilon(b, b') \in (-1/2, 1/2)$ represents the distinguishability between b and b'. We assume there exists a total ordering such that $b \succ b' \Leftrightarrow \epsilon(b, b') > 0$. We also use the notation $\epsilon_{i,j} \equiv \epsilon(b_i, b_j)$. Note that $\epsilon(b, b') = -\epsilon(b', b)$ and $\epsilon(b, b) = 0$. For ease of analysis, we also assume WLOG that the bandits are indexed in preferential order $b_1 \succ b_2 \succ \dots \succ b_K$.

Online Setting. In the online setting, algorithms are evaluated "on the fly" during every iteration. Let $(b_1^{(t)}, b_2^{(t)})$ be the bandits chosen at iteration t. Let T be the time horizon. We quantify performance using the following notion of regret,

$$R_T = \frac{1}{2} \sum_{t=1}^{T} \left(\epsilon(b_1, b_1^{(t)}) + \epsilon(b_1, b_2^{(t)}) \right). \quad (1)$$

In search applications, (1) reflects the fraction of users who would have preferred b_1 over $b_1^{(t)}$ and $b_2^{(t)}$.

PAC Setting. In the PAC setting, the goal is to confidently find a near-optimal bandit. More precisely, an (ε,δ)-PAC algorithm will find a bandit $\hat{b}$ such that $P(\epsilon(b_1, \hat{b}) > \varepsilon) \leq \delta$. Efficiency is measured via the sample complexity, i.e. the total number of comparisons required. Note that sample complexity penalizes each comparison equally, whereas the regret (1) of a comparison depends on the bandits being compared.

3.1. Modeling Assumptions

Previous work (Yue et al., 2009) relied on two properties, called stochastic triangle inequality and strong stochastic transitivity. In this paper, we assume a relaxed version of strong stochastic transitivity that more accurately characterizes real-world user preferences. Note that we only require the two properties below to be defined relative to the best bandit b_1.

Relaxed Stochastic Transitivity. For any triplet of bandits $b_1 \succ b_j \succ b_k$ and some $\gamma \geq 1$, we assume $\gamma\epsilon_{1,k} \geq \max\{\epsilon_{1,j}, \epsilon_{j,k}\}$. This can be viewed as a monotonicity or internal consistency property of user preferences. Strong stochastic transitivity, considered in (Yue et al., 2009), is the special case where $\gamma = 1$.

Stochastic Triangle Inequality. For any triplet of bandits $b_1 \succ b_j \succ b_k$, we assume $\epsilon_{1,k} \leq \epsilon_{1,j} + \epsilon_{j,k}$. This can be viewed as a diminishing returns property.[1]

To understand why relaxed stochastic transitivity is important, consider Table 1, which describes preferences elicited from pairwise interleaving experiments (Radlinski et al., 2008) using six retrieval functions in the full-text search engine[2] of ArXiv.org. We see that user preferences obey a total ordering A $\succ$ B $\succ$ ... $\succ$ F, and satisfy relaxed stochastic transitivity for $\gamma = 1.5$ (due to A, B, D) as well as stochastic triangle inequality. A good algorithm should have guarantees that degrade smoothly as γ increases.

4. Algorithm and Analysis

Our algorithm, called Beat-the-Mean, is described in Algorithm 1. The online and PAC settings require different input parameters, and those are specified in Algorithm 2 and Algorithm 3, respectively.

Beat-the-Mean proceeds in a sequence of rounds, and maintains a working set W_ℓ of active bandits during each round ℓ. For each active bandit $b_i \in W_\ell$, an empirical estimate $\hat{P}_i$ (Line 6) is maintained for how often b_i beats the *mean bandit* $\bar{b}_\ell$ of W_ℓ, where comparing b_i with $\bar{b}_\ell$ is functionally identical to comparing b_i

[1] Our results can be extended to the relaxed case where $\epsilon_{1,k} \leq \lambda(\epsilon_{1,j} + \epsilon_{j,k})$ for $\lambda \geq 1$. However, we focus on strong transitivity since it is far more easily violated in practice.

[2] http://search.arxiv.org

Algorithm 1 BEAT-THE-MEAN

1: Input: $\mathcal{B} = \{b_1, \ldots, b_K\}$, N, T, $c_{\delta,\gamma}(\cdot)$
2: $W_1 \leftarrow \{b_1, \ldots, b_K\}$ *//working set of active bandits*
3: $\ell \leftarrow 1$ *//num rounds*
4: $\forall b \in W_\ell$, $n_b \leftarrow 0$ *//num comparisons*
5: $\forall b \in W_\ell$, $w_b \leftarrow 0$ *//num wins*
6: $\forall b \in W_\ell$, $\hat{P}_b \equiv w_b/n_b$, or $1/2$ if $n_b = 0$
7: $n^* \equiv \min_{b \in W_\ell} n_b$
8: $c^* \equiv c_{\delta,\gamma}(n^*)$, or 1 if $n^* = 0$ *//confidence radius*
9: $t \leftarrow 0$ *//total number of iterations*
10: **while** $|W_\ell| > 1$ **and** $t < T$ **and** $n^* < N$ **do**
11: $b \leftarrow \text{argmin}_{b \in W_\ell} n_b$ *//break ties randomly*
12: select $b' \in W_\ell$ at random, compare b vs b'
13: if b wins, $w_b \leftarrow w_b + 1$
14: $n_b \leftarrow n_b + 1$
15: $t \leftarrow t + 1$
16: **if** $\min_{b' \in W_\ell} \hat{P}_{b'} + c^* \leq \max_{b \in W_\ell} \hat{P}_b - c^*$ **then**
17: $b' \leftarrow \text{argmin}_{b \in W_\ell} \hat{P}_b$
18: $\forall b \in W_\ell$, delete comparisons with b' from w_b, n_b
19: $W_{\ell+1} \leftarrow W_\ell \setminus \{b'\}$ *//update working set*
20: $\ell \leftarrow \ell + 1$ *//new round*
21: **end if**
22: **end while**
23: **return** $\text{argmax}_{b \in W_\ell} \hat{P}_b$

Algorithm 2 BEAT-THE-MEAN (Online)

Input $\mathcal{B} = \{b_1, \ldots, b_K\}$, γ, T
$\delta \leftarrow 1/(2TK)$
Define $c_{\delta,\gamma}(\cdot)$ using (4)
$\hat{b} \leftarrow$ BEAT-THE-MEAN($\mathcal{B}$, ∞, T, $c_{\delta,\gamma}$)

with a bandit sampled uniformly from W_ℓ (Line 12). In each iteration, a bandit with the fewest recorded comparisons is selected to compare with $\bar{b}_\ell$ (Line 11).

Whenever the empirically worst bandit b' is separated from the empirically best one by a sufficient confidence margin (Line 16), then the round ends, all recorded comparisons involving b' are removed (Line 18), and b' is removed from W_ℓ (Line 19). Afterwards, each remaining $\hat{P}_i$ is again an unbiased estimate of b_i versus the mean bandit $\bar{b}_{\ell+1}$ of the new $W_{\ell+1}$. The algorithm terminates when only one active bandit remains, or when another termination condition is met (Line 10).

Notation and terminology. We call a **round** all the contiguous comparisons until a bandit is removed. We say b_i **defeats** b_j if b_i and b_j have the highest and lowest empirical means, respectively, and that the difference is sufficiently large (Line 16). Our algorithm makes a **mistake** whenever it removes the best bandit b_1 from any W_ℓ. We will use the shorthand

$$\hat{P}_{i,j,n} \equiv \hat{P}_{i,n} - \hat{P}_{j,n}, \tag{2}$$

where $\hat{P}_{i,n}$ refers to the empirical estimate of b_i versus the mean bandit $\bar{b}_\ell$ after n comparisons (we often suppress ℓ for brevity). We call the **empirically best** and **empirically worst** bandits to be the ones with highest and lowest $\hat{P}_{i,n}$, respectively. We call the **best** and **worst** bandits in W_ℓ to be $\text{argmin}_{b_i \in W_\ell} i$ and $\text{argmax}_{b_i \in W_\ell} i$, respectively. We define the **expected performance** of any active bandit $b_i \in W_\ell$ to be

$$\mathbf{E}[\hat{P}_i] = \frac{1}{|W_\ell|} \left(\sum_{b' \in W_\ell} P(b_i > b') \right). \tag{3}$$

Algorithm 3 BEAT-THE-MEAN (PAC)

Input $\mathcal{B} = \{b_1, \ldots, b_K\}$, γ, ε, δ
Define N using (8)
Define $c_{\delta,\gamma}(\cdot)$ using (7)
$\hat{b} \leftarrow$ BEAT-THE-MEAN($\mathcal{B}$, N, ∞, $c_{\delta,\gamma}$)

For clarity of presentation, all proofs are contained in the appendix. We begin by stating two observations.

Observation 1. *Let b_k be the worst bandit in W_ℓ, and let $b_1 \in W_\ell$. Then $\mathbf{E}[\hat{P}_{1,k,n}] \geq \epsilon_{1,k}$.*

Observation 2. *Let b_k be the worst bandit in W_ℓ, and let $b_1 \in W_\ell$. Then $\forall b_j \in W_\ell : \mathbf{E}[\hat{P}_{j,k,n}] \leq 2\gamma^2 \epsilon_{1,k}$.*

Observation 1 implies a margin between the expected performance of b_1 and the worst bandit in W_ℓ. This will be used to bound the comparisons required in each round. Due to relaxed stochastic transitivity, b_1 may not have the best expected performance.[3] Observation 2 bounds the difference in expected performance between any bandit and the worst bandit in W_ℓ. This will be used to derive the appropriate confidence intervals so that $b_1 \in W_{\ell+1}$ with sufficient probability.

4.1. Online Setting

We take an "explore then exploit" approach for the online setting, similar to (Yue et al., 2009). For time horizon T, relaxed transitivity parameter γ, and bandits $\mathcal{B} = \{b_1, \ldots, b_K\}$, we use BEAT-THE-MEAN in the explore phase (see Algorithm 2). Let $\hat{b}$ denote the bandit returned by BEAT-THE-MEAN. We then enter an exploit phase by repeatedly choosing $(b_1^{(t)}, b_2^{(t)}) = (\hat{b}, \hat{b})$ until reaching T total comparisons. Comparisons in the exploit phase incur no regret assuming $\hat{b} = b_1$.

We use the following confidence interval $c(\cdot)$,

$$c_{\delta,\gamma}(n) = 3\gamma^2 \sqrt{\frac{1}{n} \log \frac{1}{\delta}}, \tag{4}$$

where $\delta = 1/(2KT)$.[4] We do not use the last remaining input N to BEAT-THE-MEAN (i.e., we set $N = \infty$); it is used only in the PAC setting.

[3] For example, the second best bandit may lose slightly to b_1 but be strongly preferred versus the other bandits.

[4] When $\gamma = 1$, we can use a tighter conf. interval (9).

We will show that BEAT-THE-MEAN correctly returns the best bandit w.p. at least $1-1/T$. Correspondingly, a suboptimal bandit is returned with probability at most $1/T$, in which case we assume maximal regret $\mathcal{O}(T)$. We can thus bound the expected regret by

$$\begin{aligned}\mathbf{E}[R_T] &\leq (1-1/T)\mathbf{E}\left[R_T^{BtM}\right] + (1/T)\mathcal{O}(T) \\ &= \mathcal{O}\left(\mathbf{E}\left[R_T^{BtM}\right] + 1\right) \end{aligned} \tag{5}$$

where R_T^{BtM} denotes the regret incurred from running BEAT-THE-MEAN. Thus the regret bound depends entirely on the regret incurred by BEAT-THE-MEAN.

Theorem 1. *For $T \geq K$, BEAT-THE-MEAN makes a mistake with probability at most $1/T$, or otherwise returns the best bandit $b_1 \in \mathcal{B}$ and accumulates online regret (1) that is bounded with high probability by*

$$\mathcal{O}\left(\sum_{\ell=1}^{K-1} \min\left\{\frac{\gamma^7}{\epsilon_\ell}, \frac{\gamma^5 \epsilon_\ell}{\epsilon_*^2}\right\} \log T\right) = \mathcal{O}\left(\frac{\gamma^7 K}{\epsilon_*} \log T\right) \tag{6}$$

where $\epsilon_\ell = \epsilon_{1,k}$ if b_k is the worst remaining bandit in round ℓ, and $\epsilon_* = \min\{\epsilon_{1,2}, \ldots, \epsilon_{1,K}\}$.

Corollary 1. *For $T \geq K$, mistake-free executions of BEAT-THE-MEAN accumulate online regret that is bounded with high probability by*

$$\mathcal{O}\left(\sum_{k=2}^{K} \frac{\gamma^8}{\epsilon_{1,k}} \log T\right).$$

Our algorithm improves on the previously proposed Interleaved Filter (IF) algorithm (Yue et al., 2009) in two ways. First, (6) applies when $\gamma > 1$,[5] whereas the bound for IF does not. Second, while (6) matches the expected regret bound for IF when $\gamma = 1$, ours is a high probability bound. Our experiments show that IF can accumulate large regret even when strong stochastic transitivity is slightly violated (e.g. $\gamma = 1.5$ as in Table 1), and has high variance when $\gamma = 1$. BEAT-THE-MEAN exhibits neither drawback.

4.2. PAC Setting

When running BEAT-THE-MEAN in the PAC setting, one of two things can happen. In the first case, the active set W_ℓ is reduced to a single bandit $\hat{b}$, in which case we will prove that $\hat{b} = b_1$ with sufficient probability. In the second case, the algorithm terminates when the number of comparisons recorded for each remaining bandit is at least N defined in (8) below, in which case we prove that every remaining bandit is within ε of b_1 with sufficient probability. It suffices to focus on the second case when analyzing sample complexity.

[5] In practice, γ is typically close to 1, making the somewhat poor dependence on γ less of a concern. This poor dependence seems primarily due to the bound of Observation 2 often being quite loose. However, it is unclear if one can do better in the general worst case.

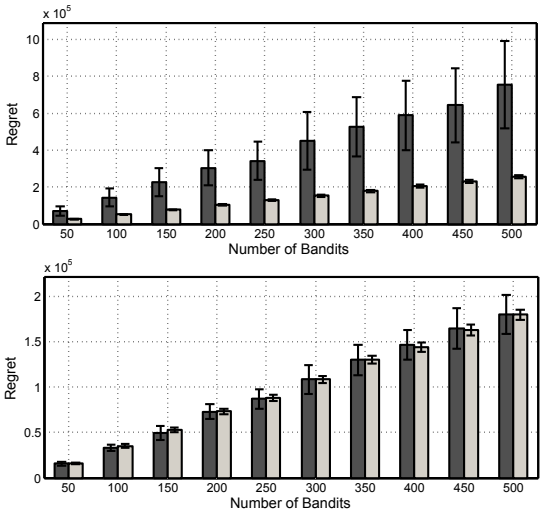

Figure 1. Comparing regret of BEAT-THE-MEAN (light) and Interleaved Filter (dark) when $\gamma = 1$. For the top graph, each $\epsilon_{i,j} = 0.1$ where $b_i \succ b_j$. For the bottom graph, each $\epsilon_{i,j} = 1/(1+\exp(\mu_j - \mu_i)) - 0.5$, where each $\mu_i \sim N(0,1)$. Error bars indicate one standard deviation.

The input parameters are described in Algorithm 3. We use the following confidence interval $c(\cdot)$,

$$c_{\delta,\gamma}(n) = 3\gamma^2 \sqrt{\frac{1}{n} \log \frac{K^3 N}{\delta}}, \tag{7}$$

where N is the *smallest positive integer* such that

$$N = \left\lceil \frac{36\gamma^6}{\varepsilon^2} \log \frac{K^3 N}{\delta} \right\rceil. \tag{8}$$

Note that there are at most $K^2 N$ total time steps, since there are at most K rounds with at most KN comparisons removed after each round.[6] We do not use the last remaining input T to BEAT-THE-MEAN (i.e., we set $T = \infty$); it is used only in the online setting.

Theorem 2. BEAT-THE-MEAN *in Algorithm 3 is an (ε, δ)-PAC algorithm with sample complexity*

$$\mathcal{O}(KN) = \mathcal{O}\left(\frac{K\gamma^6}{\varepsilon^2} \log \frac{KN}{\delta}\right).$$

5. Evaluating Online Regret

As mentioned earlier, in the online setting, the theoretical guarantees of BEAT-THE-MEAN offer two advantages over the previously proposed Interleaved Filter

[6] $K^2 N$ is a trivial bound on the sample complexity of Alg. 3. Thm. 2 gives a high probability bound of $\mathcal{O}(KN)$.

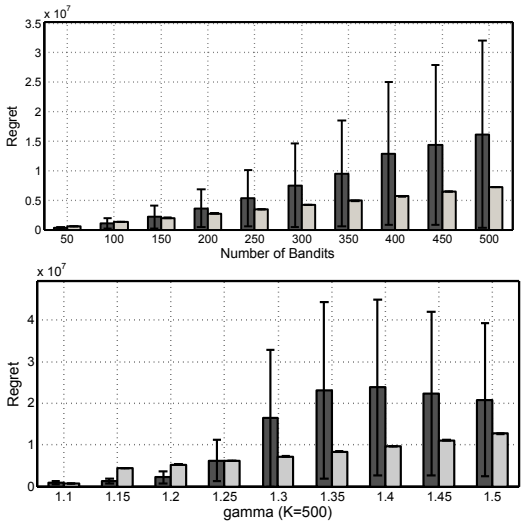

Figure 2. Comparing regret of BEAT-THE-MEAN (light) and Interleaved Filter (dark) when $\gamma > 1$. For the top graph, we fix $\gamma = 1.3$ and vary K. For the bottom graph, we fix $K = 500$ and vary γ. Error bars indicate one standard deviation.

(IF) algorithm (Yue et al., 2009) by (A) giving a high probability exploration bound versus one in expectation (and thus ensuring low variance), and (B) being provably robust to relaxations of strong transitivity ($\gamma > 1$). We now evaluate these cases empirically.

IF maintains a candidate bandit, and plays the candidate against the remaining bandits until one is confidently superior. Thus, the regret of IF heavily depends on the initial candidate, resulting in increased variance. When strong transitivity is violated ($\gamma > 1$), one can create scenarios where b_{i-1} is more likely to defeat b_i than any other bandit. Thus, IF will sift through $\mathcal{O}(K)$ candidates and suffer $\mathcal{O}(K^2)$ regret in the worst case. BEAT-THE-MEAN avoids this issue by playing every bandit against the mean bandit.

We evaluate using simulations, where we construct stochastic preferences $\epsilon_{i,j}$ that are unknown to the algorithms. Since the guarantees of BEAT-THE-MEAN and IF differ w.r.t. the number of bandits K, we fix $T = 10^{10}$ and vary K. We first evaluate Case (A), where strong transitivity holds ($\gamma = 1$). In this setting, we will use the following confidence interval,

$$c_{\delta,\gamma}(n) = \sqrt{(1/n)\log(1/\delta)}, \tag{9}$$

where $\delta = 1/(2TK)$. This is tighter than (4), leading to a more efficient algorithm that still retains the correctness guarantees when $\gamma = 1$.[7]

[7]The key observation is that cases (b) and (c) in Lemma 1 do not apply when $\gamma = 1$. The confidence interval used by IF (Yue et al., 2009) is also equivalent to (9).

We evaluate two settings for Case (A), with the results presented in Figure 1. The first setting (Figure 1 top) defines $\epsilon_{i,j} = 0.1$ for all $b_i \succ b_j$. The second setting (Figure 1 bottom) defines for each b_i a utility $\mu_i \sim N(0,1)$ drawn i.i.d. from a unit normal distribution. Stochastic preferences are defined using a logistic model, i.e. $\epsilon_{i,j} = 1/(1+\exp(\mu_j - \mu_i)) - 0.5$. For both settings, we run 100 trials and observe the average regret of both methods to increase linearly with K, but the regret of IF shows much higher variance, which matches the theoretical results.

We next evaluate Case (B), where stochastic preferences only satisfy relaxed transitivity. For any $b_i \succ b_j$, we define $\epsilon_{i,j} = 0.1\gamma$ if $1 < i = j - 1$, or $\epsilon_{i,j} = 0.1$ otherwise. Figure 2 top shows the results for $\gamma = 1.3$. We observe the similar behavior from BEAT-THE-MEAN as in Case (A),[8] but IF suffers super-linear regret (and is thus not robust) with significantly higher variance.

For completeness, Figure 2 bottom shows how regret changes as γ is varied for fixed $K = 500$. We observe a phase transition near $\gamma = 1.3$ where IF begins to suffer super-linear regret (w.r.t. K). As γ increases further, each round in IF also shortens due to each $\epsilon_{i-1,i}$ increasing, causing the regret of IF to decrease slightly (for fixed K). BEAT-THE-MEAN uses confidence intervals (4) that grow as γ increases, causing it to require more comparisons for each bandit elimination. It is possible that BEAT-THE-MEAN can still behave correctly (i.e. be mistake-free) in practice while using tighter confidence intervals (e.g. (9)).

6. Conclusion

We have presented an algorithm for the Dueling Bandits Problem with high probability exploration bounds for the online and PAC settings. The performance guarantees of our algorithm degrade gracefully as one relaxes the strong stochastic transitivity property, which is a property often violated in practice. Empirical evaluations confirm the advantages of our theoretical guarantees over previous results.

Acknowledgements. This work was funded in part by NSF Award IIS-090546.

A. Extended Analysis

Proof of Observation 1. We can write $\mathbf{E}[\hat{P}_{1,k,n}]$ (2) as

$$\mathbf{E}[\hat{P}_{1,k,n}] = \frac{1}{|W_\ell|}\left(2\epsilon_{1,k} + \sum_{b_j \in W_\ell \setminus \{b_1, b_k\}} (\epsilon_{1,j} - \epsilon_{k,j})\right).$$

[8]The regret is larger than in the analogous setting in Case (A) due to the use of wider confidence intervals.

Each b_j in the summation above satisfies $b_1 \succ b_j \succ b_k$. Thus, $\epsilon_{1,j} - \epsilon_{k,j} = \epsilon_{1,j} + \epsilon_{j,k} \geq \epsilon_{1,k}$, due to stochastic triangle inequality, implying $\mathbf{E}[\hat{P}_{1,k,n}] \geq \epsilon_{1,k}$. □

Proof of Observation 2. We focus on the non-trivial case where $b_j \neq b_k, b_1$. Combining (3) and (2) yields

$$\mathbf{E}[\hat{P}_{j,k,n}] = \frac{1}{|W_\ell|}\left(2\epsilon_{j,k} + \sum_{b_h \in W_\ell \setminus \{b_j, b_k\}} (\epsilon_{j,h} - \epsilon_{k,h})\right).$$

We know that $\epsilon_{j,k} \leq \gamma\epsilon_{1,k}$ from relaxed transitivity. For each b_h in the above summation, either (a) $b_h \succ b_j \succ b_k$, or (b) $b_j \succ b_h \succ b_k$. In case (a) we have

$$\epsilon_{j,h} - \epsilon_{k,h} = \epsilon_{j,h} + \epsilon_{h,k} \leq \epsilon_{h,k} \leq \gamma\epsilon_{1,k},$$

since $\epsilon_{j,h} \leq 0$. In case (b) we have

$$\epsilon_{j,h} + \epsilon_{h,k} \leq \gamma(\epsilon_{1,h} + \epsilon_{1,k}) \leq 2\gamma \max\{\epsilon_{1,h}, \epsilon_{1,k}\} \leq 2\gamma^2\epsilon_{1,k}.$$

This implies that $\mathbf{E}[\hat{P}_{j,k,n}] \leq 2\gamma^2\epsilon_{1,k}$. □

A.1. Online Setting

We assume here that Beat-the-Mean is run according to Alg. 2. We define $c_n \equiv c_{\delta,\gamma}(n)$ using (4).

Lemma 1. *For $\delta = 1/(2TK)$, and assuming that $b_1 \in W_\ell$, the probability of $b_i \in W_\ell \setminus \{b_1\}$ defeating b_1 at the end of round ℓ (i.e., a mistake) is at most $1/(TK^2)$.*

Proof. Having b_i defeat b_1 requires that for some n, $\hat{P}_{1,n} + c_n < \hat{P}_{i,n} - c_n$, and also that b_1 is the *first* bandit to be defeated in the round. Since at any time, all remaining bandits have the same confidence interval size, then $\hat{P}_{1,n}$ must have the lowest empirical mean. In particular, this requires $\hat{P}_{1,n} \leq \hat{P}_{k,n}$, where b_k is the worst bandit in W_ℓ. We will show that, for any n, the probability of making a mistake is at most $2\delta^2 < 1/(T^2K^2)$. Thus, by the union bound, the probability of b_i mistakenly defeating b_1 for any $n \leq T$ is at most $2T\delta^2 < 1/(TK^2)$. We consider three sufficient cases:

(a) $\mathbf{E}[\hat{P}_{i,n}] \leq \mathbf{E}[\hat{P}_{1,n}]$

(b) $\mathbf{E}[\hat{P}_{i,n}] > \mathbf{E}[\hat{P}_{1,n}]$ and $n < \frac{4}{\epsilon_{1,k}^2}\log(1/\delta)$

(c) $\mathbf{E}[\hat{P}_{i,n}] > \mathbf{E}[\hat{P}_{1,n}]$ and $n \geq \frac{4}{\epsilon_{1,k}^2}\log(1/\delta)$

In case (a) applying Hoeffding's inequality yields

$$P(\hat{P}_{1,n} + c_n < \hat{P}_{i,n} - c_n) \leq 2\delta^4 < 2\delta^2.$$

In cases (b) and (c), we have $b_i \neq b_k$ since Observation 1 implies $\mathbf{E}[\hat{P}_{1,n}] \geq \mathbf{E}[\hat{P}_{k,n}]$. In case (b) we have $c_n > (3/2)\gamma^2\epsilon_{1,k}$, which implies via Hoeffding's inequality,

$$\begin{aligned}
&P(\hat{P}_{i,n} - c_n > \hat{P}_{1,n} + c_n) \\
&\quad\leq P(\hat{P}_{i,n} - c_n > \hat{P}_{k,n} + c_n) \qquad (10)\\
&\quad= P(\hat{P}_{i,k,n} - \mathbf{E}[\hat{P}_{i,k,n}] > 2c_n - \mathbf{E}[\hat{P}_{i,k,n}]) \\
&\quad\leq P(\hat{P}_{i,k,n} - \mathbf{E}[\hat{P}_{i,k,n}] > 2c_n - 2\gamma^2\epsilon_{1,k}) \qquad (11)\\
&\quad\leq P(\hat{P}_{i,k,n} - \mathbf{E}[\hat{P}_{i,k,n}] > (2/3)c_n) \\
&\quad\leq \exp\left(-2\gamma^4\log(1/\delta)\right) \leq \delta^2 < 2\delta^2
\end{aligned}$$

where (10) follows from $\forall j : \epsilon_{1,j} \geq \epsilon_{k,j}$, and (11) follows from Observation 2. In case (c) we know that

$$\begin{aligned}
P(\hat{P}_{k,n} \geq \hat{P}_{1,n}) &= P(\hat{P}_{k,1,n} - \mathbf{E}[\hat{P}_{k,1,n}] \geq -E[\hat{P}_{k,1,n}]) \\
&\leq P(\hat{P}_{k,1,n} - \mathbf{E}[\hat{P}_{k,1,n}] \geq \epsilon_{1,k}) \qquad (12)\\
&\leq \exp\left(-n\epsilon_{1,k}^2/2\right) = \delta^2 < 2\delta^2
\end{aligned}$$

□

Lemma 2. Beat-the-Mean *makes a mistake with probability at most $1/T$.*

Proof. By Lemma 1, the probability $b_j \in W_\ell \setminus \{b_1\}$ defeats b_1 is at most $1/(TK^2)$. There are at most K active bandits in any round and at most K rounds. Applying the union bound proves the lemma. □

Lemma 3. *Let $\delta = 1/(TK)$, $T \geq K$ and assume $b_1 \in W_\ell$. If b_k is the worst bandit in W_ℓ, then the number of comparisons each $b \in W_\ell$ needs to accumulate before some bandit being removed (and thus ending the round) is with high probability bounded by*

$$\mathcal{O}\left(\frac{\gamma^4}{\epsilon_{1,k}^2}\log(TK)\right) = \mathcal{O}\left(\frac{\gamma^4}{\epsilon_{1,k}^2}\log T\right).$$

Proof. It suffices to bound the comparisons n required to remove b_k. We will show that for any $d \geq 1$, there exists an m depending only on d such that

$$P\left(n \geq \frac{m\gamma^4}{\epsilon_{1,k}^2}\log(TK)\right) \leq \min\left\{K^{-d}, T^{-d}\right\}$$

for all K and T sufficiently large. We will focus on the sufficient condition of b_1 defeating b_k: if at any t we have $\hat{P}_{1,t} - c_t > \hat{P}_{k,t} + c_t$, then b_k is removed from W_ℓ. It follows that for any t, if $n > t$, then $\hat{P}_{1,t} - c_t \leq \hat{P}_{k,t} + c_t$, and so $P(n > t) \leq P(\hat{P}_{1,t} - c_t \leq \hat{P}_{k,t} + c_t)$.

Note from Observation 1 that $\mathbf{E}[\hat{P}_{1,k,t}] \geq \epsilon_{1,k}$. Thus,

$$\begin{aligned}
&P(\hat{P}_{1,t} - c_t \leq \hat{P}_{k,t} + c_t) \\
&\quad= P(\mathbf{E}[\hat{P}_{1,k,t}] - \hat{P}_{1,k,t} \geq \mathbf{E}[\hat{P}_{1,k,t}] - 2c_t) \\
&\quad\leq P(\mathbf{E}[\hat{P}_{1,k,t}] - \hat{P}_{1,k,t} \geq \epsilon_{1,k} - 2c_t) \qquad (13)
\end{aligned}$$

For any $m \geq 18$ and $t \geq \lceil 8m\gamma^4 \log(2TK)/\epsilon_{1,k}^2 \rceil$, we have $c_t \leq \gamma\epsilon_{1,k}/4$, and so applying Hoeffding's inequality for this m and t shows that (13) is bounded by

$$\leq P(|\hat{P}_{1,k,t} - \mathbf{E}[\hat{P}_{1,k,t}]| \geq \epsilon_{1,k}/2) \leq 2\exp(-t\epsilon_{1,k}^2/8).$$

Since $t \geq 8m\gamma^4 \log(2TK)/\epsilon_{1,k}^2$ by assumption, we have $t\epsilon_{1,k}^2/8 \geq m\log(2TK)$, and so

$$2\exp(-t\epsilon_{1,k}^2/8) \leq 2\exp(-m\log(2TK)) = 1/(TK)^m,$$

which is bounded by K^{-m} and T^{-m}. We finally note that for $T \geq K$, $\mathcal{O}(\log(2TK)) = \mathcal{O}(\log T)$. □

Lemma 4. *Assume $b_1 \in W_{\ell'}$ for $\ell' \leq \ell$. Then the number of comparisons removed at the end of round ℓ is bounded with high probability by*

$$\mathcal{O}\left(\min\left\{\frac{\gamma^4}{\epsilon_*^2}, \frac{\gamma^6}{\epsilon_\ell^2}\right\}\log T\right),$$

where $\epsilon_\ell = \epsilon_{1,K}$ if b_k is the worst bandit in W_ℓ, and $\epsilon_* = \min\{\epsilon_{1,2}, \ldots, \epsilon_{1,K}\}$.

Proof. Let b_k be the worst bandit in W_ℓ, and let $b_j \in W_\ell$ denote the bandit removed at the end of round ℓ. By Lemma 3, the total number of comparisons that b_j accumulates is bounded with high probability by

$$\mathcal{O}\left(\frac{\gamma^4}{\epsilon_{1,k}^2}\log T\right) = \mathcal{O}\left(\frac{\gamma^4}{\epsilon_*^2}\log T\right). \tag{14}$$

Some bandits may have accumulated more comparisons than (14), since the number of remaining comparisons from previous rounds may exceed (14). However, Lemma 3 implies the number of remaining comparisons is at most

$$\mathcal{O}\left(\frac{|W_\ell|\gamma^4}{\epsilon_{1,k'}^2}\log T\right),$$

where $\epsilon_{1,k'} = \min_{b_k \succ b_q} \epsilon_{1,q} \geq \epsilon_{1,k}/\gamma$. We can bound the number of comparisons accumulated at the end of round ℓ by $\mathcal{O}(|W_\ell| D_{\ell,\gamma,T})$, where

$$D_{\ell,\gamma,T} \equiv \min\left\{\frac{\gamma^4}{\epsilon_*^2}, \frac{\gamma^6}{\epsilon_\ell^2}\right\}\log T, \tag{15}$$

and $\epsilon_\ell = \epsilon_{1,k}$. In expectation

$$\frac{1 + (|W_\ell| - 1)/|W_\ell|}{|W_\ell|} < \frac{2}{|W_\ell|} \tag{16}$$

fraction of the comparisons will be removed at the end of round ℓ (those that involve b_j). Applying Hoeffding's inequality yields the number of removed comparisons is with high probability $\mathcal{O}(D_{\ell,\gamma,T})$. □

Proof of Theorem 1. Lemma 2 bounds the mistake probability by $1/T$. We focus here on the case where Beat-the-Mean is mistake-free. We will show that, at the end of each round ℓ, the regret incurred from the removed comparisons is bounded with high probability by $\mathcal{O}(\gamma\epsilon_\ell D_{\ell,\gamma,T})$ for $D_{\ell,\gamma,T}$ defined in (15). Since this happens $K-1$ times (once per round) and accounts for the regret of every comparison, then the total accumulated regret of Beat-the-Mean will be bounded with high probability by

$$\mathcal{O}\left(\sum_{\ell=1}^{K-1}\gamma\epsilon_\ell D_{\ell,\gamma,T}\right) = \mathcal{O}\left(\sum_{\ell=1}^{K-1}\min\left\{\frac{\gamma^5\epsilon_\ell}{\epsilon_*^2}, \frac{\gamma^7}{\epsilon_\ell}\right\}\log T\right).$$

By Lemma 4, the number of removed comparisons in round ℓ is at most $\mathcal{O}(D_{\ell,\gamma,T})$. The incurred regret of a comparison between any b_i and the removed bandit b_j is $(\epsilon_{1,i}+\epsilon_{1,j})/2 \leq \gamma\epsilon_\ell$, which follows from relaxed transitivity. Thus, the regret incurred from all removed comparisons of round ℓ is at most $\mathcal{O}(\gamma\epsilon_\ell D_{\ell,\gamma,T})$. □

Proof of Corollary 1. We know from Theorem 1 that the regret assigned to the removed bandit in each round ℓ is $\mathcal{O}((\gamma^7/\epsilon_\ell)\log T)$, where $\epsilon_\ell = \epsilon_{1,k}$ if b_k is the worst bandit in W_ℓ. By the pigeonhole principle $b_{K+1-\ell} \succ b_k$, since in round ℓ there are $K+1-\ell$ bandits. By relaxed transitivity, we have $\epsilon_{1,K+1-\ell} \leq \gamma\epsilon_{1,k}$. The desired result naturally follows. □

A.2. PAC Setting

We assume here that Beat-the-Mean is run according to Alg. 3. We define $c_n \equiv c_{\delta,\gamma}(n)$ using (7).

Lemma 5. *Assuming that $b_1 \in W_\ell$, the probability of $b_i \in W_\ell \setminus \{b_1\}$ defeating b_1 at the end of round ℓ (i.e., a mistake) is at most δ/K^3.*

Proof. (Sketch). This proof is structurally identical to Lemma 1, except using a different confidence interval. We consider three analogous cases:

(a) $\mathbf{E}[\hat{P}_{i,n}] < \mathbf{E}[\hat{P}_{1,n}]$

(b) $\mathbf{E}[\hat{P}_{i,n}] \geq \mathbf{E}[\hat{P}_{1,n}]$ and $n < \frac{4}{\epsilon_{1,k}^2}\log(K^3N/\delta)$

(c) $\mathbf{E}[\hat{P}_{i,n}] \geq \mathbf{E}[\hat{P}_{1,n}]$ and $n \geq \frac{4}{\epsilon_{1,k}^2}\log(K^3N/\delta)$

In case (a) applying Hoeffding's inequality yields

$$P(\hat{P}_{1,n} + c_n < \hat{P}_{i,n} - c_n) \leq 2(\delta/(K^3N))^4 < \delta/(K^5N).$$

In case (b), following (10) and (11), we have

$$\begin{aligned} &P(\hat{P}_{i,n} - c_n > \hat{P}_{1,n} + c_n) \\ &\quad \leq P(\hat{P}_{i,k,n} - \mathbf{E}[\hat{P}_{i,k,n}] > (2/3)c_n) \\ &\quad \leq \exp\left(-2\gamma^4\log(1/\delta)\right) \leq (\delta/(K^3N))^2 < \delta/(K^5N). \end{aligned}$$

In case (c), following (12) and using $K \geq 2$, we have

$$\begin{aligned} P(\hat{P}_{1,n} \leq \hat{P}_{k,n}) &\leq 2\exp\left(-n\epsilon_{i,k}^2/2\right) \\ &\leq 2\exp\left(-2\log(K^3N/\delta)\right) \\ &= 2(\delta/(K^3N))^2 < \delta/(K^5N). \end{aligned}$$

So the probability of each failure case is at most $\delta/(K^5N)$. There are at most K^2N time steps, so applying the union bound proves the claim. □

Lemma 6. Beat-the-Mean *makes a mistake with probability at most* δ/K.

The proof of Lemma 6 is exactly the same as Lemma 2, except leveraging Lemma 5 instead of Lemma 1.

Lemma 7. *If a mistake-free execution of* Beat-the-Mean *terminates due to* $n^* = N$ *in round* ℓ*, then* $P(\exists b_j \in W_\ell : \epsilon_{1,j} > \varepsilon) \leq \delta/K$.

Proof. Suppose Beat-the-Mean terminates due to $n^* = N$ after round ℓ and t total comparisons. For any $b_i \in W_\ell$, we know via Hoeffding's inequality that

$$P(|\hat{P}_{i,N} - \mathbf{E}[\hat{P}_{i,N}]| \geq c_N) \leq 2\exp\left(-18\gamma^4\log(K^3N/\delta)\right),$$

which is at most $\delta/(K^4N)$. There are at most K^2N comparisons, so taking the union bound over all t and b_i yields δ/K. So with probability at least $1 - \delta/K$, each $\hat{P}_i$ is within c_N of its expectation when Beat-the-Mean terminates. Using (8), this implies

$$\forall b_j \in W_\ell,\ \mathbf{E}[\hat{P}_1] - \mathbf{E}[\hat{P}_j] \leq 2c_N \leq \varepsilon/\gamma.$$

In particular, for the worst bandit b_k in W_ℓ we have

$$\varepsilon/\gamma \geq \mathbf{E}[\hat{P}_1] - \mathbf{E}[\hat{P}_k] \geq \epsilon_{1,k}, \tag{17}$$

which follows from Observation 1. Since $\gamma\epsilon_{1,k} \geq \epsilon_{1,j}$ for any $b_j \in W_\ell$, then (17) implies $\varepsilon \geq \epsilon_{1,j}$. □

Proof of Theorem 2. We first analyze correctness. We consider two sufficient failure cases:

(a) Beat-the-Mean makes a mistake

(b) Beat-the-Mean is mistake-free and there exists an active bandit b_j upon termination where $\epsilon_{1,j} > \varepsilon$

Lemma 6 implies that the probability of (a) is at most δ/K. If Beat-the-Mean terminates due to $n^* = N$, then Lemma 7 implies that the probability of (b) is also at most δ/K. By the union bound the total probability of (a) or (b) is bounded by $2\delta/K \leq \delta$, since $K \geq 2$.

We now analyze the sample complexity. Let L denote the round when the termination condition $n^* = N$ is satisfied ($L = K - 1$ if the condition was never satisfied). For rounds $1, \ldots, L-1$, the maximum number of unremoved comparisons accumulated by any bandit is N. Using the same argument from (16), the number of comparisons removed after each round is then with high probability $\mathcal{O}(N)$. In round L, each of the $K - L + 1$ remaining bandits accumulate at most N comparisons, implying that the total number of comparisons made is bounded by

$$\mathcal{O}\left((K - L + 1)N + (L - 1)N\right) = \mathcal{O}\left(\frac{K\gamma^6}{\varepsilon^2}\log\frac{KN}{\delta}\right).$$

□

References

Agarwal, Alekh, Dekel, Ofer, and Xiao, Lin. Optimal algorithms for online convex optimization with multi-point bandit feedback. In *Conference on Learning Theory (COLT)*, 2010.

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

Even-Dar, Eyal, Mannor, Shie, and Mansour, Yishay. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research (JMLR)*, 7:1079–1105, 2006.

Feige, Uriel, Raghavan, Prabhakar, Peleg, David, and Upfal, Eli. Computing with noisy information. *SIAM Journal on Computing*, 23(5), 1994.

Kalyanakrishnan, Shivaram and Stone, Peter. Efficient selection of multiple bandit arms: Theory and practice. In *International Conference on Machine Learning (ICML)*, 2010.

Lai, T. L. and Robbins, Herbert. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4–22, 1985.

Mannor, Shie and Tsitsiklis, John N. The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research (JMLR)*, 5: 623–648, 2004.

Radlinski, Filip and Joachims, Thorsten. Active exploration for learning rankings from clickthrough data. In *ACM Conference on Knowledge Discovery and Data Mining (KDD)*, 2007.

Radlinski, Filip, Kurup, Madhu, and Joachims, Thorsten. How does clickthrough data reflect retrieval quality? In *ACM Conference on Information and Knowledge Management (CIKM)*, 2008.

Yue, Yisong and Joachims, Thorsten. Interactively optimizing information retrieval systems as a dueling bandits problem. In *International Conference on Machine Learning (ICML)*, 2009.

Yue, Yisong, Broder, Josef, Kleinberg, Robert, and Joachims, Thorsten. The k-armed dueling bandits problem. In *Conference on Learning Theory (COLT)*, 2009.