The K-armed Dueling Bandits Problem

Yisong Yue
Dept. of Computer Science
Cornell University
Ithaca, NY 14853
yyue@cs.cornell.edu

Josef Broder
Center for Applied Math
Cornell University
Ithaca, NY 14853
jbroder@cam.cornell.edu

Robert Kleinberg
Dept. of Computer Science
Cornell University
Ithaca, NY 14853
rdk@cs.cornell.edu

Thorsten Joachims
Dept. of Computer Science
Cornell University
Ithaca, NY 14853
tj@cs.cornell.edu

Abstract

We study a partial-information online-learning problem where actions are restricted to noisy comparisons between pairs of strategies (also known as bandits). In contrast to conventional approaches that require the absolute reward of the chosen strategy to be quantifiable and observable, our setting assumes only that (noisy) binary feedback about the relative reward of two chosen strategies is available. This type of relative feedback is particularly appropriate in applications where absolute rewards have no natural scale or are difficult to measure (e.g., user-perceived quality of a set of retrieval results, taste of food, product attractiveness), but where pairwise comparisons are easy to make. We propose a novel regret formulation in this setting, as well as present an algorithm that achieves (almost) information-theoretically optimal regret bounds (up to a constant factor).

1 Introduction

In partial information online learning problems (also known as bandit problems) [Rob52], an algorithm must choose, in each of T consecutive iterations, one of K possible bandits (strategies). For conventional bandit problems, in every iteration, each bandit receives a real-valued payoff in $[0, 1]$, initially unkown to the algorithm. The algorithm then chooses one bandit and receives (and thus observes) the associated payoff. No other payoffs are observed. The goal then is to maximize the total payoff (i.e., the sum of payoffs over all iterations).

The conventional setting assumes that observations perfectly reflect (or are unbiased estimates of) the received payoffs. In many applications, however, such observations may be unavailable or unreliable. Consider, for example, applications in sensory testing or information retrieval, where the payoff is the goodness of taste or the user-perceived quality of a retrieval result. While it is difficult to elicit payoffs on an absolute scale in such applications, one can reliably obtain relative judgments of payoff (i.e. “A tastes better than B”, or “ranking A is better than ranking B”). In fact, user behavior can often be modeled as maximizing payoff, so that such relative comparison statements can be derived from observable user behavior. For example, to elicit whether a search-engine user prefers ranking r_1 over r_2 for a given query, Radlinski et al. [RKJ08] showed how to present an interleaved ranking of r_1 and r_2 so that clicks indicate which of the two is preferred by the user. This ready availability of pairwise comparison feedback in applications where absolute payoffs are difficult to observe motivates our learning framework.

Given a collection of K bandits (e.g., retrieval functions), we wish to find a sequence of noisy comparisons that has low regret. We call this the *K-armed Dueling Bandits Problem*, which can also be viewed as a regret-minimization version of the classical problem of finding the maximum element of a set using noisy comparisons [FRPU94]. A canonical application example is an intranet-search system that is installed for a new customer. Among K built-in retrieval functions, the search engine needs to select the one that provides the best results on this collection, with pairwise feedback coming from clicks in the interleaved rankings [RKJ08]. Since the search engine incurs regret whenever it presents the results from a suboptimal retrieval function, it aims to identify suboptimal retrieval functions as quickly as possible. More generally, the Dueling Bandits Problem arises naturally in many applications where a system must adapt interactively to specific user bases, and where pairwise comparisons are easier to elicit than absolute payoffs.

One important issue is formulating an appropriate notion of regret. Since we are concerned with maximizing user utility (or satisfaction), but utility is not directly quantifiable in our pairwise-comparison model, a natural question to ask is whether users, at each iteration, would have prefered another bandit over the ones chosen by our algorithm. This leads directly to our regret formulation (described in Section 3), which measures regret based on the (initially unknown) probability that the best bandit b^* would win a comparison with the chosen bandits at each iteration. One can alternatively view this as the fraction of users who would have prefered b^* over the bandits chosen by our algorithm.

Our solution follows an “explore then exploit” approach, where we will bound expected regret by the regret incurred while running the exploration algorithm. We will present two exploration algorithms in Section 4, which we call Interleaved Filter 1 and Interleaved Filter 2. Interleaved Filter 1 incurs regret that, with high probability, is within a logarithmic factor of the information-theoretic optimum. Interleaved Filter 2 uses an interesting extension to achieve expected regret that is within a constant factor of the information-theoretic optimum. We will prove the matching lower bound in Section 5.

An interesting feature of our Interleaved Filter algorithms is that, unlike previous search algorithms based on noisy comparisons, e.g., [FRPU94], the number of experiments devoted to each bandit during the exploration phase is highly non-uniform: of the K bandits, there is a small subset of bandits ($\mathcal{O}(\log K)$ of them in expectation) who each participate in $\mathcal{O}(K)$ comparisons, while the remaining bandits only participate in $\mathcal{O}(\log K)$ comparisons in expectation. In Section 5 we provide insight about why existing methods suffer high regret in our setting. Thus, our results provide theoretical support for Langford's observation [Lan08] about a qualitative difference between algorithms for supervised learning and those for learning from partial observations: in the supervised setting, "holistic information is often better," whereas in the setting of partial observations it is often better to select a few points and observe them many times while giving scant attention to other points.

2 Related Work

Regret-minimizing algorithms for multi-armed bandit problems and their generalizations have been intensively studied for many years, both in the stochastic [LR85] and non-stochastic [ACBFS02] cases. The vast literature on this topic includes algorithms whose regret is within a constant factor of the information-theoretic lower bound in the stochastic case [ACBF02] and within a $O(\sqrt{\log n})$ factor of the best such lower bound in the non-stochastic case [ACBFS02]. Our use of upper confidence bounds in designing algorithms for the dueling bandits problem is prefigured by their use in the multi-armed bandit algorithms that appear in [Aue03, ACBF02, LR85].

Upper confidence bounds are also central to the design of multi-armed bandit problems in the PAC setting [EDMM06, MT04], where the algorithm's objective is to identify an arm that is ε-optimal with probability at least $1 - \delta$. Our work adopts a very different feedback model (pairwise comparisons rather than direct observation of payoffs) and a different objective (regret minimization rather than the PAC objective) but there are clear similarities between our IF1 and IF2 algorithms and the Successive Elimination and Median Eliminiation algorithms developed for the PAC setting in [EDMM06]. There are also some clear differences between the algorithms: these are discussed in Section 5.1.

The difficulty of the dueling bandits problem stems from the fact that the algorithm has no way of directly observing the costs of the actions it chooses. It is an example of a *partial monitoring problem*, a class of regret-minimization problems defined in [CBLS06], in which an algorithm (the "forecaster") chooses actions and then observes feedback signals that depend on the actions chosen by the forecaster and by an unseen opponent (the "environment"). This pair of actions also determines a loss, which is not revealed to the forecaster but is used in defining the forecaster's regret. Under the crucial assumption that the feedback matrix has high enough rank that its row space spans the row space of the loss matrix (which is required in order to allow for a Hannan consistent forecaster) the results of [CBLS06] show that there is a forecaster whose regret is bounded by $O(T^{2/3})$ against a non-stochastic (adversarial) environment, and that there exist partial monitoring problems for which this bound can not be improved. Our dueling bandits problem is a special case of the partial monitoring problem, our environment is stochastic rather than adversarial, and thus our regret bound exhibits much better (i.e., logarithmic) dependence on T.

Banditized online learning problems based on absolute rewards (of individual actions) have been previously studied in the context of web advertising [PACJ07, LZ07]. In that setting, clear explicit feedback is available in the form of (expected) revenue. We study settings where such absolute measures are unavailable or unreliable.

Our work is also closely related to the literature on computing with noisy comparison operations [AGHB$^+$94, BOH08, FRPU94, KK07], in particular the design of tournaments to identify the maximum element in an ordered set, given access to noisy comparators. All of these papers assume unit cost per comparison, whereas we charge a different cost for each comparison depending on the pair of elements being compared. In the unit-cost-per-comparison model, and assuming that every comparison has ϵ probability of error regardless of the pair of elements being compared, Feige et al. [FRPU94] presented sequential and parallel algorithms that achieve the information-theoretically optimal expected cost (up to constant factors) for many basic problems such as sorting, searching, and selecting the maximum. The upper bound for noisy binary search has been improved in a very recent paper [BOH08] that achieves the information-theoretic optimum up to a $1+o(1)$ factor. When the probability of error depends on the pair of elements being compared (as in our dueling bandits problem), Adler et al. [AGHB$^+$94] and Karp and Kleinberg [KK07] present algorithms that achieve the information-theoretic optimum (up to constant factors) for the problem of selecting the maximum and for binary search, respectively. Our results can be seen as extending this line of work to the setting of regret minimization. It is worth noting that the most efficient algorithms for selecting the maximum in the model of noisy comparisons with unit cost per comparison [AGHB$^+$94, FRPU94] are not suitable in the regret minimization setting considered here, because they devote undue effort to comparing elements that are far from the maximum. This point is discussed further in Section 5.1.

Yue and Joachims [YJ09] simultaneously studied a continuous version of the Dueling Bandits Problem, where bandits (e.g., retrieval functions) are characterized using a compact parameter space. For that setting, they proposed a gradient descent algorithm which achieves sublinear regret (with respect to the time horizon). In many applications, it may be infeasible or undesirable to interactively explore such a large space of bandits. For instance, in intranet search one might reasonably "cover" the space of plausible retrieval functions with a small number of hand-crafted retrieval functions. In such cases, selecting the best of K well-engineered solutions would be much more efficient than searching a possibly huge space of real-valued parameters.

Learning based on pairwise comparisons is well studied in the (off-line) supervised learning setting called learning to rank. Typically, a preference function is first learned using a set of i.i.d. training examples, and subsequent predictions are made to minimize the number of mis-ranked pairs (e.g., [CSS99]). Most prior work assume access to a training set with absolute labels (e.g., of relevance or utility) on individual examples, with pairwise preferences generated using inputs with labels from different ordinal classes (e.g., [HGO99, FISS03, Joa05, BBB$^+$07, LS07, AM08]). In the case where there are exactly two label classes, this becomes the so-called bipartite ranking problem [BBB$^+$07, AM08], which is a more general version of learning to optimize ROC-Area [HGO99, Joa05, LS07]. All known prior work assume that the training data defines (or is used to generate) preferences between individual examples, rather than between hypothesis functions (which is our setting).

3 The Dueling Bandits Problem

We propose a new online optimization problem, called the K-armed Dueling Bandits Problem, where the goal is to find the best among K bandits $\mathcal{B} = \{b_1, \ldots, b_K\}$. Each iteration comprises of a noisy comparison (a duel) between two bandits (possibly the same bandit with itself). We assume that the outcomes of these noisy comparisons are independent random variables and that the probability of b winning a comparison with b' is stationary over time. We write this probability as $P(b > b') = \epsilon(b, b') + 1/2$, where $\epsilon(b, b') \in (-1/2, 1/2)$ is a measure of the distinguishability between b and b'. We assume that there exists a total ordering on $\mathcal{B}$ such that $b \succ b'$ implies $\epsilon(b, b') > 0$. We will also use the notation $\epsilon_{i,j} \equiv \epsilon(b_i, b_j)$.

Let $(b_1^{(t)}, b_2^{(t)})$ be the bandits chosen at iteration t, and let b^* be the overall best bandit. We define **strong regret** based on comparing the chosen bandits with b^*,

$$R_T = \sum_{t=1}^{T} \left(\epsilon(b^*, b_1^{(t)}) + \epsilon(b^*, b_2^{(t)}) \right), \tag{1}$$

where T is the time horizon. We also define **weak regret**,

$$\tilde{R}_T = \sum_{t=1}^{T} \min\{\epsilon(b^*, b_1^{(t)}), \epsilon(b^*, b_2^{(t)})\}, \tag{2}$$

which only compares $\hat{b}$ against the better of $b_1^{(t)}$ and $b_2^{(t)}$. One can regard strong regret as the fraction of users who would have preferred the best bandit over the chosen ones in each iteration[1]. (More precisely, it corresponds to the fraction of users who prefer the best bandit to a uniformly-random member of the pair of bandits chosen, in the case of strong regret, or to the better of the two bandits chosen, in the case of weak regret.) We will present algorithms which achieve identical regret bounds for both formulations (up to constant factors) by assuming a property called stochastic triangle inequality, which is described in the next section.

3.1 Assumptions

We impose additional structure to the probabilistic comparisons. First, we assume **strong stochastic transitivity**, which requires that any triplet of bandits $b_i \succ b_j \succ b_k$ satisfies

$$\epsilon_{i,k} \geq \max\{\epsilon_{i,j}, \epsilon_{j,k}\}. \tag{3}$$

Algorithm 1 Explore Then Exploit Solution

```
Input: $T$, $\mathcal{B} = \{b_1, \ldots, b_K\}$, $EXPLORE$
$(\hat{b}, \hat{T}) \leftarrow EXPLORE(T, \mathcal{B})$
for $t = \hat{T} + 1, \ldots, T$ do
    compare $\hat{b}$ and $\hat{b}$
end for
```

This assumption provides a monotonicity constraint on possible probability values.

We also assume **stochastic triangle inequality**, which requires any triplet of bandits $b_i \succ b_j \succ b_k$ to satisfy

$$\epsilon_{i,k} \leq \epsilon_{i,j} + \epsilon_{j,k}. \tag{4}$$

Stochastic triangle inequality captures the condition that the probability of a bandit winning (or losing) a comparison will exhibit diminishing returns as it becomes increasingly superior (or inferior) to the competing bandit[2].

We briefly describe two common generative models which satisfy these two assumptions. The first is the logistic or Bradley-Terry model, where each bandit b_i is assigned a positive real value μ_i. Probabilistic comparisons are made using

$$P(b_i > b_j) = \frac{\mu_i}{\mu_i + \mu_j}.$$

The second is a Gaussian model, where each bandit is associated with a random variable X_i that has a Gaussian distribution with mean μ_i and variance 1. Probabilistic comparisons are made using

$$P(b_i > b_j) = P(X_i - X_j > 0),$$

where $X_i - X_j \sim N(\mu_i - \mu_j, 2)$. It is straightforward to check that both models satisfy strong stochastic transitivity and stochastic triangle inequality.

4 Algorithm & Analysis

Our solution, which is described in Algorithm 1, follows an explore then exploit approach. For a given time horizon T and a set of K bandits $\mathcal{B} = \{b_1, \ldots, b_K\}$, an exploration algorithm (denoted generically as EXPLORE) is used to find the best bandit b^*. EXPLORE returns both its solution $\hat{b}$ as well as the total number of iterations $\hat{T}$ for which it ran (it is possible that $\hat{T} > T$). Should $\hat{T} < T$, we enter an exploit phase by repeatedly choosing $(b_1^{(t)}, b_2^{(t)}) = (\hat{b}, \hat{b})$, which incurs no additional regret assuming EXPLORE correctly found the best bandit ($\hat{b} = b^*$). In the case where $\hat{T} > T$, then the regret incurred from running EXPLORE still bounds our regret formulations (which only measure regret up to T), so our analysis in this section will still hold.

We will consider two versions of our proposed exploration algorithm, which we call Interleaved Filter 1 (IF1) and Interleaved Filter 2 (IF2). We will show that both algorithms (which we refer to generically as IF) correctly return the best bandit with probability at least $1 - 1/T$. Correspondingly, a suboptimal bandit is returned with probability at most $1/T$,

[1] In the search setting, users experience an interleaving, or mixing, of results from both retrieval functions to be compared.

[2] Our analysis also applies for a relaxed version where $\epsilon_{i,k} \leq \gamma(\epsilon_{i,j} + \epsilon_{j,k})$ for finite $\gamma > 0$.

in which case we assume maximal regret $\mathcal{O}(T)$. We can thus bound the expected regret by

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

where R_T^{IF} denotes the regret incurred from running Interleaved Filter. Thus the regret bound depends entirely on the regret incurred by Interleaved Filter.

The two IF algorithms are described in Algorithm 2 and Algorithm 3, respectively. IF2 achieves an expected regret bound which matches the information-theoretic lower bound (up to constant factors) presented in Section 5, whereas IF1 matches with high probability the lower bound up to a log factor. We first examine IF1 due to its ease of analysis. We then analyze IF2, which builds upon IF1 to achieve the information-theoretic optimum.

In both versions, IF maintains a candidate bandit $\hat{b}$ and simulates simultaneously comparing $\hat{b}$ with all other remaining bandits via round robin scheduling (i.e., interleaving). Any bandit that is empirically inferior to $\hat{b}$ with $1 - \delta$ confidence is removed (we will describe later how to choose δ). When some bandit b' is empirically superior to $\hat{b}$ with $1 - \delta$ confidence, then $\hat{b}$ is removed and b' becomes the new candidate $\hat{b} \leftarrow b'$. IF2 contains an additional step where all empirically inferior bandits (even if lacking $1 - \delta$ confidence) are removed (called pruning – see lines 16-18 in Algorithm 3). This process repeats until only one bandit remains. Assuming IF has not made any mistakes, then it will return the best bandit $\hat{b} = b^*$.

Terminology. Interleaved Filter makes a "**mistake**" if it draws a false conclusion regarding a pair of bandits. A mistake occurs when an inferior bandit is determined with $1 - \delta$ confidence to be the superior one. We call the additional step of IF2 (lines 16-18 in Algorithm 3) "**pruning**". We define a "**match**" to be all the comparisons Interleaved Filter makes between two bandits, and a "**round**" to be all the matches played by one candidate $\hat{b}$. We always refer to $\log x$ as the natural log, $\ln x$, whenever the distinction is necessary.

In our analysis, we assume WLOG that the bandits in $\mathcal{B}$ are sorted in preferential order $b_1 \succ \ldots \succ b_K$. Then for $T \geq K$, we will show in Theorem 1 that running IF1 incurs, with high probability, regret bounded by

$$R_T^{IF1} = \mathcal{O}\left(\frac{K \log K}{\epsilon_{1,2}} \log T\right).$$

Note that $\epsilon_{1,2} = P(b_1 \succ b_2) - 1/2$ is the distinguishability between the two best bandits. Due to strong stochastic transitivity, $\epsilon_{1,2}$ lower bounds the distinguishability between the best bandit and any other bandit. We will also show in Theorem 2 that running IF2 incurs expected regret bounded by

$$\mathbf{E}\left[R_T^{IF2}\right] = \mathcal{O}\left(\frac{K}{\epsilon_{1,2}} \log T\right),$$

which matches the information-theoretic lower bound (up to constant factors) described in Section 5.

Analysis Approach. Our analysis follows three phases. We first bound the regret incurred for any match. Then for both IF1 and IF2, we show that the mistake probability is at most $1/T$. We finally bound the matches played by IF1 and IF2 to arrive at our final regret bounds.

Algorithm 2 Interleaved Filter 1 (IF1)

```
Input: $T$, $\mathcal{B} = \{b_1, \ldots, b_K\}$
$\delta \leftarrow 1/(TK^2)$
Choose $\hat{b} \in \mathcal{B}$ randomly
$W \leftarrow \{b_1, \ldots, b_K\} \setminus \{\hat{b}\}$
$\forall b \in W$, maintain estimate $\hat{P}_{\hat{b},b}$ of $P(\hat{b} > b)$
$\forall b \in W$, maintain $1 - \delta$ confidence interval $\hat{C}_{\hat{b},b}$ of $\hat{P}_{\hat{b},b}$
while $W \neq \emptyset$ do
    for $b \in W$ do
        compare $\hat{b}$ and $b$
        update $\hat{P}_{\hat{b},b}$, $\hat{C}_{\hat{b},b}$
    end for
    while $\exists b \in W$ s.t. $\left(\hat{P}_{\hat{b},b} > 1/2 \wedge 1/2 \notin \hat{C}_{\hat{b},b}\right)$ do
        $W \leftarrow W \setminus \{b\}$
    end while
    if $\exists b' \in W$ s.t. $\left(\hat{P}_{\hat{b},b'} < 1/2 \wedge 1/2 \notin \hat{C}_{\hat{b},b'}\right)$ then
        $\hat{b} \leftarrow b'$, $W \leftarrow W \setminus \{b'\}$   //new round
        $\forall b \in W$, reset $\hat{P}_{\hat{b},b}$ and $\hat{C}_{\hat{b},b}$
    end if
end while
$\hat{T} \leftarrow$ Total Comparisons Made
return $(\hat{b}, \hat{T})$
```

4.1 Confidence Intervals

In a match between b_i and b_j, IF maintains a number $\hat{P}_{i,j}$, the empirical estimate of $P(b_i \succ b_j)$ after t comparisons[3]. For ease of notation, we drop the subscripts (b_i, b_j), and use $\hat{P}_t$, which emphasizes the dependence on the number of comparisons. IF similarly maintains a confidence interval

$$\hat{C}_t = (\hat{P}_t - c_t, \hat{P}_t + c_t),$$

where $c_t = \sqrt{\log(1/\delta)/t}$. We justify the construction of these confidence intervals in the following lemma.

Lemma 1. *For $\delta = 1/(TK^2)$, the number of comparisons in a match between b_i and b_j is with high probability at most*

$$\mathcal{O}\left(\frac{1}{\epsilon_{i,j}^2} \log(TK)\right).$$

Moreover, the winner is identified correctly with probability at least $1 - \delta$, provided $\delta \in (0, 1/2]$.

Proof. We first argue correctness. Fix the number of comparisons t, and note that $\mathbf{E}[\hat{P}_t] = 1/2 + \epsilon_{i,j}$. This tells us $P(1/2 + \epsilon_{i,j} \notin \hat{C}_t)$ is bounded above by the probability that $\hat{P}_t$ deviates from its expected value by at least c_t. An application of Hoeffding's inequality [Hoe63] shows that this probability is bounded above by

$$2\exp(-2tc_t^2) = 2\exp(-2\log(1/\delta)) \leq \delta$$

[3] In other words, $\hat{P}_{i,j}$ is the fraction of these t comparisons in which b_i was the winner.

for $\delta \le 1/2$. Thus for every t, we know with confidence at least $1-\delta$ that $1/2+\epsilon_{i,j} \in \hat{C}_t$. It follows from this fact and the stopping conditions of the match that the winner is correctly identified with probability at least $1-\delta$.

To bound the number of comparisons n in a match, it suffices to prove that for any $d \ge 1$, there exists an m depending only on d such that

$$P\left(n \ge \frac{m}{\epsilon_{i,j}^2}\log(TK)\right) \le K^{-d}$$

for all K sufficiently large. Assume without loss of generality that $\epsilon_{i,j} > 0$, and define the event $\mathcal{E}_t = \{\hat{P}_t - c_t < 1/2\}$. Note that $\mathcal{E}_t$ is a necessary condition for the match to continue after t comparisons, so for any $t \in \mathbb{N}$,

$$P(n > t) \le P(\mathcal{E}_t).$$

For $m \ge 4$ and $t = \lceil m \log(TK^2)/\epsilon_{i,j}^2 \rceil$, we have $c_t \le \epsilon_{i,j}/2$. Applying Hoeffding's inequality for this t shows

$$\begin{aligned}
P(\mathcal{E}_t) &= P(\hat{P}_t - (1/2+\epsilon_{i,j}) < c_t - \epsilon_{i,j}) \\
&= P(\mathbf{E}[\hat{P}_t] - \hat{P}_t > \epsilon_{i,j} - c_t) \\
&\le P(|\hat{P}_t - \mathbf{E}[\hat{P}_t]| > \epsilon_{i,j}/2) \\
&\le 2\exp(-t\epsilon_{i,j}^2/2) \\
&\le 2\exp(-m\log(TK^2)) \\
&= 2/(TK^2)^m.
\end{aligned}$$

Taking $m = \max\{4, d\}$ proves the lemma for all $K \ge 2$. □

4.2 Regret per Match

We now bound the accumulated regret of each match.

Lemma 2. *Assuming b_1 has not been removed and $T \ge K$, then with high probability the accumulated weak regret and also (assuming stochastic triangle inequality) strong regret from any match is at most*

$$\mathcal{O}\left(\frac{1}{\epsilon_{1,2}}\log T\right).$$

Proof. Suppose the candidate bandit $\hat{b} = b_j$ is playing a match against b_i. Since all matches within a round are played simultaneously, then by Lemma 1, any match played by b_j contains at most

$$\mathcal{O}\left(\frac{1}{\epsilon_{1,j}^2}\log(TK)\right) \le \mathcal{O}\left(\frac{1}{\epsilon_{1,2}^2}\log(TK)\right)$$

comparisons, where the inequality follows from strong stochastic transitivity. Note that $\min\{\epsilon_{1,j}, \epsilon_{1,i}\} \le \epsilon_{1,j}$. Then the accumulated weak regret (2) is bounded by

$$\begin{aligned}
\epsilon_{1,j}\mathcal{O}\left(\frac{1}{\epsilon_{1,j}^2}\log(TK)\right) &= \mathcal{O}\left(\frac{1}{\epsilon_{1,j}}\log(TK)\right) \\
&\le \mathcal{O}\left(\frac{1}{\epsilon_{1,2}}\log(TK)\right) \\
&= \mathcal{O}\left(\frac{1}{\epsilon_{1,2}}\log T\right) \qquad (6)
\end{aligned}$$

where (6) holds since $\log(TK) \le \log(T^2) = 2\log T$. We now bound the accumulated strong regret (1) by leveraging stochastic triangle inequality. Each comparison incurs $\epsilon_{1,j} + \epsilon_{1,i}$ regret. We now consider three cases.

Case 1: Suppose $b_i \succ b_j$. Then $\epsilon_{1,j} + \epsilon_{1,i} \le 2\epsilon_{1,j}$, and the accumulated strong regret of the match is bounded by

$$2\epsilon_{1,j}\mathcal{O}\left(\frac{1}{\epsilon_{1,j}^2}\log(TK)\right) \quad \le \mathcal{O}\left(\tfrac{1}{\epsilon_{1,2}}\log(TK)\right)$$

Case 2: Suppose $b_j \succ b_i$ and $\epsilon_{j,i} \le \epsilon_{1,j}$. Then

$$\begin{aligned}
\epsilon_{1,j} + \epsilon_{1,i} &\le \epsilon_{1,j} + \epsilon_{1,j} + \epsilon_{j,i} \\
&\le 3\epsilon_{1,j}
\end{aligned}$$

and the accumulated strong regret is bounded by

$$\begin{aligned}
3\epsilon_{1,j}\mathcal{O}\left(\frac{1}{\epsilon_{1,j}^2}\log(TK)\right) &= \mathcal{O}\left(\tfrac{1}{\epsilon_{1,j}}\log(TK)\right) \\
&\le \mathcal{O}\left(\tfrac{1}{\epsilon_{1,2}}\log(TK)\right)
\end{aligned}$$

Case 3: Suppose $b_j \succ b_i$ and $\epsilon_{j,i} > \epsilon_{1,j}$. Then we can also use Lemma 1 to bound with high probability the number of comparisons by

$$\mathcal{O}\left(\frac{1}{\epsilon_{j,i}^2}\log(TK)\right).$$

The accumulated strong regret is then bounded by

$$\begin{aligned}
3\epsilon_{j,i}\mathcal{O}\left(\frac{1}{\epsilon_{j,i}^2}\log(TK)\right) &= \mathcal{O}\left(\tfrac{1}{\epsilon_{j,i}}\log(TK)\right) \\
&\le \mathcal{O}\left(\tfrac{1}{\epsilon_{1,j}}\log(TK)\right) \\
&\le \mathcal{O}\left(\tfrac{1}{\epsilon_{1,2}}\log(TK)\right)
\end{aligned}$$

Like in the analysis for weak regret (6), we finally note that

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

□

In the next two sections, we will bound the mistake probability and total matches played by IF1 and IF2, respectively.

4.3 Regret Bound for Interleaved Filter 1

We first state our main regret bound for Interleaved Filter 1.

Theorem 1. *Running Algorithm 1 with $\mathcal{B} = \{b_1, \ldots, b_K\}$, time horizon T ($T \ge K$), and IF1 incurs expected regret (both weak and strong) bounded by*

$$\mathbf{E}[R_T] \le \mathcal{O}\left(\mathbf{E}\left[R_T^{IF1}\right]\right) = \mathcal{O}\left(\frac{K\log K}{\epsilon_{1,2}}\log T\right).$$

The proof follows immediately from combining Lemma 3, Lemma 5, Lemma 2 and (5). We begin by analyzing the probability of IF1 making a mistake.

Lemma 3. *For $\delta \leq 1/(TK^2)$, IF1 makes a mistake with probability at most $1/T$*

Proof. By Lemma 1, the probability that IF1 makes a mistake in any given match is at most $1/(TK^2)$. Since K^2 is a trivial upper bound on the number of matches, applying the union bound over all matches proves the lemma. □

We assume for the remainder of this section that IF1 is mistake-free, since the cost of making a mistake is considered in (5), and we are interested here in bounding R_T^{IF1}.

Random Walk Model. We can model the sequence of candidate bandits as a random walk. Let each bandit be a node on a graph, where b_j ($j > 1$) transitions to b_i for $1 \leq i < j$ with some probability. The best bandit b_1 is an absorbing node. Due to strong stochastic transitivity, the probability of b_j transitioning to b_i is at least the probability of transitioning to b_h (for $h > i$). We will consider the worst case where b_j transitions to each of $b_1, \ldots, b_{j-1}$ with equal probability. We can thus bound the number of rounds required by IF1 (and also IF2) by analyzing the length of a random walk from b_K to b_1. We will prove that this random walk requires $\mathcal{O}(\log K)$ steps with high probability.

Let X_i ($1 \leq i < K$) be an indicator random variable corresponding to whether a random walk starting at b_K visits b_i in the Random Walk Model.

Lemma 4. *For X_i as defined above with $1 \leq i < K$,*

$$P(X_i = 1) = \frac{1}{i},$$

and furthermore, for all $S \subseteq \{X_1, \ldots, X_{K-1}\}$,

$$P(S) = \prod_{X_i \in S} P(X_i), \tag{7}$$

meaning $X_1, \ldots, X_{K-1}$ are mutually independent.

Proof. We can rewrite (7) as

$$P(S) = \prod_{X_i \in S} P(X_i | S_i),$$

where $S_i = \{X_j \in S | j > i\}$.

We first consider $S = \{X_1, \ldots, X_{K-1}\}$. For the factor on X_i, denote with j the smallest index in S_i with $X_j = 1$ in the condition. Then

$$\begin{aligned}&P(X_i = 1 | X_{i+1}, ..., X_{K-1}) \\ &\quad = P(X_i = 1 | X_{i+1} = 0, ..., X_{j-1} = 0, X_j = 1) = \frac{1}{i}\end{aligned}$$

since the walk moved to one of the first i nodes with uniform probability independent of j. Since $\forall j > i : P(X_i = 1 | X_j = 1) = \frac{1}{i}$, this implies $P(X_i = 1) = \frac{1}{i}$. So we can conclude

$$P(X_1, \ldots, X_{K-1}) = \prod_{i=1}^{K-1} P(X_i).$$

Now consider arbitrary S. We use $\sum_{S^c}$ to indicate summing over the joint states of all X_i variables not in S. We can write $P(S)$ as

$$\begin{aligned} P(S) &= \sum_{S^c} P(X_1, \ldots, X_{K-1}) \\ &= \sum_{S^c} \prod_{i=1}^{K-1} P(X_i) \\ &= \prod_{X_i \in S} P(X_i) \left(\sum_{S^c} \prod_{X_i \in S^c} P(X_i) \right) \\ &= \prod_{X_i \in S} P(X_i). \end{aligned}$$

This proves mutual independence (7).

□

We can express the number of steps taken by a random walk from b_K to b_1 in the Random Walk Model as $S_K = 1 + \sum_{i=1}^{K-1} X_i$. Lemma 4 implies that

$$E[S_K] = 1 + \sum_{i=1}^{K-1} E[X_i] = 1 + H_{K-1} \approx \log K,$$

where H_i is the harmonic sum. We now show that $S_K = \mathcal{O}(\log K)$ with high probability.

Lemma 5. *Assuming IF1 is mistake-free, then it runs for $\mathcal{O}(\log K)$ rounds with high probability.*

Proof. Consider the Random Walk Model. It suffices to show that for any $d \geq 1$, there exists a m depending only on d such that

$$\forall K \geq 1 : \quad P(S_K > m \log K) \leq \frac{1}{K^d}. \tag{8}$$

Using the Chernoff bound [MR95], we know that for any $m > 1$,

$$\begin{aligned} P(S_K > m(1 + H_{K-1})) \quad &\leq \left(\tfrac{e^{m-1}}{m^m}\right)^{1+H_{K-1}} \\ &\leq \left(\tfrac{e^{m-1}}{m^m}\right)^{1+\log K} \\ &= (eK)^{m-1-m\log m} \end{aligned} \tag{9}$$

(9) is true since

$$\log K \leq H_{K-1} < \log K + 1.155$$

for all $K \geq 1$ (where 1.155 is approximately twice Euler's constant). We require this bound to be at most $1/K^d$. Solving

$$(eK)^{m-1-m\log m} \leq K^{-d}$$

yields $m \geq e^d$. The Chernoff bound applies for all $K \geq 0$. So for any $d \geq 1$, we can choose $m = e^d$ to satisfy (8). □

Algorithm 3 Interleaved Filter 2 (IF2)

```
Input: $T$, $\mathcal{B} = \{b_1, \dots, b_K\}$
$\delta \leftarrow 1/(TK^2)$
Choose $\hat{b} \in \mathcal{B}$ randomly
$W \leftarrow \{b_1, \dots, b_K\} \setminus \{\hat{b}\}$
$\forall b \in W$, maintain estimate $\hat{P}_{\hat{b},b}$ of $P(\hat{b} > b)$
$\forall b \in W$, maintain $1-\delta$ confidence interval $\hat{C}_{\hat{b},b}$ of $\hat{P}_{\hat{b},b}$
while $W \neq \emptyset$ do
    for $b \in W$ do
        compare $\hat{b}$ and $b$
        update $\hat{P}_{\hat{b},b}$, $\hat{C}_{\hat{b},b}$
    end for
    while $\exists b \in W$ s.t. $\left(\hat{P}_{\hat{b},b} > 1/2 \wedge 1/2 \notin \hat{C}_{\hat{b},b}\right)$ do
        $W \leftarrow W \setminus \{b\}$
    end while
    if $\exists b' \in W$ s.t. $\left(\hat{P}_{\hat{b},b'} < 1/2 \wedge 1/2 \notin \hat{C}_{\hat{b},b'}\right)$ then
        while $\exists b \in W$ s.t. $\hat{P}_{\hat{b},b} > 1/2$ do
            $W \leftarrow W \setminus \{b\}$    //pruning
        end while
        $\hat{b} \leftarrow b',\ W \leftarrow W \setminus \{b'\}$    //new round
        $\forall b \in W$, reset $\hat{P}_{\hat{b},b}$ and $\hat{C}_{\hat{b},b}$
    end if
end while
$\hat{T} \leftarrow$ Total Comparisons Made
return $(\hat{b}, \hat{T})$
```

4.4 Regret Bound for Interleaved Filter 2

We first state our main regret bound for Interleaved Filter 2.

Theorem 2. *Running Algorithm 1 with $\mathcal{B} = \{b_1, \dots, b_K\}$, time horizon T ($T \geq K$), and IF2 incurs expected regret (both weak and strong) bounded by*

$$\mathbf{E}[R_T] \leq \mathcal{O}\left(\mathbf{E}\left[R_T^{IF2}\right]\right) = \mathcal{O}\left(\frac{K}{\epsilon_{1,2}} \log T\right).$$

The proof follows immediately from combining Lemma 6, Lemma 7, Lemma 2 and (5). IF2 improves upon IF1 by removing all empirically inferior bandits whenever the candidate is defeated, which we call pruning. We begin by analyzing the pruning technique.

Theorem 3. *For $\delta \leq 1/(TK^2)$, when the incumbent bandit $\hat{b}$ is defeated with $1-\delta$ confidence by some other bandit b', then for all bandits b'' found to be empirically inferior (but lacking $1-\delta$ confidence) to $\hat{b}$, we can conclude that b' is superior to b'' with $1-\delta$ confidence.*

Proof. Suppose a round just ended, where the incumbent candidate $\hat{b} = b_j$ has been defeated by b_i. We treat i as a random variable, since any remaining bandit could, in principle, have defeated b_j. Suppose also that bandit b_k was found empirically inferior to b_j (but lacking $1-\delta$ confidence).

Define n to be the number of comparisons made for each match in this round, and let $S_{i,j}$ denote the number of comparisons b_i won versus b_j. By definition of our confidence intervals (see Section 4.1), we know that

$$S_{i,j} - \frac{n}{2} > \sqrt{n \log\left(\frac{1}{\delta}\right)}, \qquad (10)$$

since that is the stopping condition for the round.

Let $A_{n,i,k}$ denote the event that b_i is the bandit that defeats b_j in n comparisons, and $b_k \succ b_i$, but IF2 mistakenly concludes $b_i \succ b_k$. Then for any b_k, it suffices to prove

$$P\left(\bigcup_{i,n} A_{n,i,k}\right) \leq \delta.$$

By taking the union bound, we have

$$\begin{aligned} P\left(\bigcup_{n,i} A_{n,i,k}\right) &\leq \sum_{n,i} P(A_{n,i,k}) \\ &= \sum_{n,i} P(n,i) P(B_{n,i,k}|n,i) \\ &= \mathbf{E}_{n,i}[P(B_{n,i,k}|n,i)] \qquad (11) \end{aligned}$$

where $P(n,i)$ is the joint probability that the round lasts for n comparisons per match with b_i winning the round. $B_{n,i,k}$ denotes the event that, *conditioned* on n and b_i winning, IF2 mistakenly concludes $b_i \succ b_k$. We will show that for all n and b_i, we can reject $B_{n,i,k}$ with confidence $1-\delta$ and thus can bound (11) by δ.

Suppose that $P(b_i > b_j) = \alpha$. Then under $B_{n,i,k}$, we know from strong stochastic transitivity that $P(b_k > b_j) \geq \alpha$, and therefore $P(b_j > b_k) \leq 1-\alpha$ and $\mathbf{E}[S_{i,j} + S_{j,k}] \leq n$. Combining $S_{j,k} \geq n/2$ with (10) yields

$$S_{i,j} + S_{j,k} - n > \sqrt{n \log\left(\frac{1}{\delta}\right)}. \qquad (12)$$

Then we can consider

$$P\left(S_{i,j} + S_{j,k} - n > \sqrt{n \log\left(\frac{1}{\delta}\right)}\right). \qquad (13)$$

We will analyze the worst case when $P(b_j > b_k) = 1-\alpha$. Deviating from this worst case will only make Hoeffding's inequality on (13) tighter. In this worst case, for any α we have $\mathbf{E}[S_{i,j} + S_{j,k}] = n$. Using Hoeffding's inequality [Hoe63], we have

$$\begin{aligned} P\left(S_{i,j} + S_{j,k} - n > \sqrt{n \log\left(\tfrac{1}{\delta}\right)}\right) &\leq \exp\left\{\tfrac{-2n \log(1/\delta)}{2n}\right\} \\ &= \delta \end{aligned}$$

We can thus reject $B_{n,i,k}$ with confidence at least $1-\delta$. □

Lemma 6. *For $\delta \leq 1/(TK^2)$, IF2 makes a mistake with probability at most $1/T$*

Proof. In round r, suppose b_i is the incumbent, and suppose b_i is not defeated. Then every match in this round is played to completion, so by Lemma 1, the probability that IF2 makes a mistake in any single match is at most $1/(TK^2)$. An application of the union bound shows that IF2 makes a mistake in this round with probability at most $1/(TK)$. On the other hand, suppose that in round r, b_i is defeated by b_k. In this case, IF2 makes a mistake only if it discards a bandit b_j when in fact $b_j \succ b_k$. By Theorem 3, we know the probability of this event is bounded above by $1/(TK^2)$, so the probability that IF2 makes a mistake in this round is at most $1/(TK)$. Taking the union bound over all rounds proves the lemma. □

For the remainder of this section, we analyze the behavior of IF2 when it is mistake-free. We will show that, in expectation, IF2 plays $O(K)$ matches and thus incurs expected regret bounded by

$$\mathcal{O}\left(\frac{K}{\epsilon_{1,2}} \log T\right).$$

Lemma 7. *Assuming IF2 is mistake free, then it plays $\mathcal{O}(K)$ matches in expectation.*

Proof. We leverage the Random Walk Model defined in Section 4.3 in order to provide a worst case analysis. Let B_j denote a random variable counting the number of matches played by b_j when it is *not* the candidate (to avoid double-counting). We can write B_j as

$$B_j = A_j + G_j,$$

where A_j indicates the number of matches played by b_j against b_i for $i > j$ (when the candidate was inferior to b_j), and G_j indicates the number of matches played by b_j against b_i for $i < j$ (when the candidate was superior to b_j). We can thus bound the expected number of matches played by

$$\sum_{j=1}^{K-1} \mathbf{E}[B_j] = \sum_{j=1}^{K-1} \mathbf{E}[A_j] + \mathbf{E}[G_j]. \quad (14)$$

By Lemma 4, we can write $\mathbf{E}[A_j]$ as

$$\mathbf{E}[A_j] = 1 + \sum_{i=j+1}^{K-1} \frac{1}{i} = 1 + H_{K-1} - H_i,$$

where H_i is the harmonic sum.

We now analyze $\mathbf{E}[G_j]$. We assume the worst case that $\mathbf{E}[G_j]$ does not lose a match (with $1-\delta$ confidence) to any superior candidate b_i before the match concludes (b_i is defeated) unless $b_i = b_1$. We can thus bound $\mathbf{E}[G_j]$ using the probability that b_j is pruned at the conclusion of each round. Let $\mathcal{E}_{j,t}$ denote the event that b_j is pruned after the tth round in which the candidate bandit is superior to b_j, conditioned on not being pruned in the first $t-1$ such rounds. Define $G_{j,t}$ to indicate the number of matches beyond the first $t-1$ played by b_j against a superior candidate, conditioned on playing at least $t-1$ such matches. We can write $\mathbf{E}[G_{j,t}]$ as

$$\mathbf{E}[G_{j,t}] = 1 + P(\mathcal{E}_{j,t}^c)\mathbf{E}[G_{j,t+1}],$$

and thus

$$\mathbf{E}[G_j] \leq \mathbf{E}[G_{j,1}] \leq 1 + P(\mathcal{E}_{j,1}^c)\mathbf{E}[G_{j,2}]. \quad (15)$$

We know that $P(\mathcal{E}_{j,t}^c) \leq 1/2$ for all $j \neq 1$ and t. From Lemma 5, we know that $\mathbf{E}[G_{j,t}] \leq \mathcal{O}(K \log K)$ and is thus finite. Hence, we can bound (15) by the infinite geometric series $1 + 1/2 + 1/4 + \ldots = 2$.

We can thus write (14) as

$$\begin{aligned}\sum_{j=1}^{K-1} \mathbf{E}[A_j] + \mathbf{E}[G_j] &\leq \sum_{j=1}^{K-1} (1 + H_{K-1} - H_j) + 2(K-1) \\ &= \sum_{j=1}^{K-1} (j-1)\frac{1}{j} + 3(K-1) = \mathcal{O}(K).\end{aligned}$$

□

5 Lower Bounds

We now show that the bound in Theorem 2 is information theoretically optimal up to constant factors. The proof is similar to the lower bound proof for the standard stochastic multi-armed bandit problem. However, since we make a number of assumptions not present in the standard case (such as a total ordering of $\mathcal{B}$), we present a simple self-contained lower bound argument, rather than a reduction from the standard case.

Theorem 4. *Any algorithm ϕ for the dueling bandits problem has*

$$R_T^\phi = \Omega\left(\frac{K}{\epsilon} \log T\right),$$

where $\epsilon = \min_{b \neq b^*} P(b^*, b)$.

The proof is motivated by Lemma 5 of [KNMS08]. Fix $\epsilon > 0$ and define the following family of problem instances. In instance $\tilde{q}_j$, let b_j be the best bandit, and order the remaining bandits by their indices. Let $P(b_i \succ b_k) = \epsilon$ whenever $b_i \succ b_k$. Note that these are valid problem instances, i.e. they satisfy (3), (4), and the assumptions in Section 3.

Let q_j be the distribution on T-step histories induced by instance $\tilde{q}_j$. Let $n_{j,T}$ be the number of matches involving bandit b_j scheduled by ϕ up to time T. Using these instances, we will prove a lemma from which Theorem 4 will follow.

Lemma 8. *Let ϕ be an algorithm for the dueling bandits problem such that*

$$R_T^\phi = o(T^a)$$

for all $a > 0$. Then for all j,

$$\mathbf{E}_{q_1}[n_{j,T}] = \Omega\left(\frac{\log T}{\epsilon^2}\right).$$

Proof. Fix $j \neq 1$ and $0 < a < 1/2$. Define the event $\mathcal{E}_j = \{n_{j,T} < \log(T)/\epsilon^2\}$. If $q_1(\mathcal{E}_j) < 1/3$, then

$$\mathbf{E}_{q_1}[n_{j,T}] \geq q_1(\mathcal{E}_j^c)(\log(T)/\epsilon^2) = \Omega\left(\frac{\log T}{\epsilon^2}\right).$$

So suppose now that $q_1(\mathcal{E}_j) \geq 1/3$. Under q_j, the algorithm incurs regret ϵ for every comparison involving a bandit $b \neq b_j$. This fact together with the assumption on ϕ imply that $\mathbf{E}_{q_j}[T - n_{j,T}] = o(T^a)$. Using this fact and Markov's inequality, we have

$$\begin{aligned} q_j(\mathcal{E}_j) &= q_j(\{T - n_{j,T} > T - \log(T)/\epsilon^2\}) \\ &\leq \frac{\mathbf{E}_{q_j}[T - n_{j,T}]}{T - \log(T)/\epsilon^2} = o(T^{a-1}). \end{aligned}$$

In [KK07], Karp and Kleinberg prove that for any event $\mathcal{E}$ and distributions p, q with $p(\mathcal{E}) \geq 1/3$ and $q(\mathcal{E}) < 1/3$,

$$KL(p;q) \geq \frac{1}{3} \ln \left(\frac{1}{3q(\mathcal{E})} \right) - \frac{1}{e}.$$

Applying this lemma with the event $\mathcal{E}_j$, we have

$$\begin{aligned} KL(q_1;q_j) &\geq \frac{1}{3} \ln \left(\frac{1}{3o(T^{a-1})} \right) - \frac{1}{e} \\ &= \Omega(\log T) \qquad (16) \end{aligned}$$

On the other hand, by the chain rule for KL divergence [CT99], we have

$$\begin{aligned} KL(q_1;q_j) &= \mathbf{E}_{q_1}[n_{j,T}] KL(1/2+\epsilon; 1/2-\epsilon) \\ &\leq 16\epsilon^2 \mathbf{E}_{q_1}[n_{j,T}] \qquad (17) \end{aligned}$$

Combining (16) and (17) proves the lemma. □

Proof of Theorem 4. Let ϕ be any algorithm for the dueling bandits problem. If ϕ does not satisfy the hypothesis of Lemma 8, the theorem holds trivially. Otherwise, on the problem instance specified by q_1, ϕ incurs regret at least ϵ every time it plays a match involving $b_j \neq b_1$. It follows from Lemma 8 that

$$R_T^{\phi} \geq \sum_{j \neq 1} \epsilon \mathbf{E}_{q_1}[n_{j,T}] = \Omega \left(\frac{K}{\epsilon} \log T \right).$$

□

5.1 Discussion of Related Work

Algorithms for finding maximal elements in a noisy information model are discussed in [FRPU94]. That paper describes a tournament-style algorithm that returns the best of K elements with probability $1-\delta$ in $O(K \log(1/\delta)/\epsilon^2)$ comparisons, where ϵ is the minimum margin of victory of one element over an inferior one. This is achieved by arranging the elements in a binary tree and running a series of mini-tournaments, in which a parent and its two children compete until a winner can be identified with high confidence. Winning nodes are promoted to the parent position, and lower levels of the tree are pruned to reduce the total number of comparisons. The maximal element eventually reaches the root of the tree with high probability.

Such a tournament could incur very high regret in our framework. Consider a mini-tournament involving three suboptimal but barely distinguishable elements (e.g. $P(b^* \succ b_{i,j,k}) \approx 1$, but $P(b_i \succ b_j) = 1/2 + \gamma$ for $\gamma << 1$). This tournament would require $\Omega(1/\gamma^2)$ comparisons to determine the best element, but each comparison would contribute $\Omega(1)$ to the total regret. Since γ can be arbitrarily small compared to $\epsilon^* = \epsilon_{1,2}$, this yields a regret bound that can be arbitrarily worse than the above lower bound. In general, algorithms that achieve low regret in our model must avoid such situations, and must discard suboptimal bandits after as few comparisons as possible. This heuristic motivates the interleaved structure proposed in our algorithms, which allows for good control over the number of matches involving suboptimal bandits.

This discussion also sheds light on the reasons our algorithms for the dueling bandits problem differ from algorithms that achieve optimal or near-optimal sample complexity bounds for multi-armed bandit problems in the PAC setting [EDMM06]. As mentioned in Section 2, there are striking similarities between our IF1 algorithm and the Successive Elimination algorithm from [EDMM06] as well as similarities between our IF2 algorithm and the Median Elimination algorithm from [EDMM06]. However, as explained in the preceding paragraph, in our setting all of the highly suboptimal arms (those contributing significantly more than ϵ regret per sample) must be eliminated quickly (before sampling more than ϵ^{-2} times). In the Successive/Median Elimination algorithms, every arm is sampled at least ϵ^{-2} times. The need to eliminate highly suboptimal arms quickly is specific to the regret minimization setting and exerts a strong influence on the design of the algorithm; in particular, it motivates the interleaved structure as explained above. This design choice prompts another feature of our algorithms that distinguishes them from the Successive/Median Elimination algorithms, namely the choice of an “incumbent” arm in each phase that participates in many more samples than the other arms. The algorithms for the PAC setting [EDMM06] distribute the sampling load evenly among all arms participating in a phase.

6 Conclusion

We have proposed a novel framework for partial information online learning in which feedback is derived from pairwise comparisons, rather than absolute measures of utility. We have defined a natural notion of regret for this problem, and designed algorithms that are information theoretically optimal for this performance measure. Our results extend previous work on computing in noisy information models, and is motivated by practical considerations from information retrieval applications. Future directions include finding other reasonable notions of regret in this framework (e.g., via contextualization [LZ07]), and designing algorithms that achieve low-regret when the set of bandits is very large (a special case of this is addressed in [YJ09]).

Acknowledgments

The work is funded by NSF Award IIS-0812091. The first author is also supported by a Microsoft Research Graduate Fellowship and a Yahoo! Key Technical Challenges Grant. The third author is supported by NSF Awards CCF-0643934 and CCF-0729102, an Alfred P. Sloan Foundation Fellowship, and a Microsoft Research New Faculty Fellowship.

The authors also thank Filip Radlinski and the anonymous reviewers for their helpful suggestions.

References

[ACBF02] Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multi-armed bandit problem. *Machine Learning*, 47(2):235–256, 2002.

[ACBFS02] Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

[AGHB$^+$94] Micah Adler, Peter Gemmell, Mor Harchol-Balter, Richard Karp, and Claire Kenyon. Selection in the presence of noise: The design of playoff systems. In *ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 1994.

[AM08] Nir Ailon and Mehryar Mohri. An efficient reduction of ranking to classification. In *Conference on Learning Theory (COLT)*, 2008.

[Aue03] Peter Auer. Using confidence bounds for exploitation-exploration trade. *Journal of Machine Learning Research (JMLR)*, 3:397–422, 2003.

[BBB$^+$07] Maria-Florina Balcan, Nikhil Bansal, Alina Beygelzimer, Don Coppersmith, John Langford, and Gregory Sorkin. Robust reductions from ranking to classification. In *Conference on Learning Theory (COLT)*, 2007.

[BOH08] Michael Ben-Or and Avinatan Hassidim. The bayesian learner is optimal for noisy binary search (and pretty good for quantum as well). In *IEEE Symposium on Foundations of Computer Science (FOCS)*, 2008.

[CBLS06] Nicolò Cesa-Bianchi, Gábor Lugosi, and Gilles Stoltz. Regret minimization under partial monitoring. *Mathematics of Operations Research*, 31(3):562–580, 2006.

[CSS99] William Cohen, Robert Schapire, and Yoram Singer. Learning to order things. *Journal of Artificial Intelligence Research (JAIR)*, 10:243–270, 1999.

[CT99] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. J. Wiley, 1999.

[EDMM06] Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research (JMLR)*, 7:1079–1105, 2006.

[FISS03] Yoav Freund, Raj Iyer, Robert Schapire, and Yoram Singer. An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research (JMLR)*, 4:933–969, 2003.

[FRPU94] Uriel Feige, Prabhakar Raghavan, David Peleg, and Eli Upfal. Computing with noisy information. *SIAM Journal on Computing*, 23(5), 1994.

[HGO99] Ralf Herbrich, Thore Graepel, and Klaus Obermayer. Support vector learning for ordinal regression. In *International Conference on Artificial Neural Networks (ICANN)*, 1999.

[Hoe63] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

[Joa05] Thorsten Joachims. A support vector method for multivariate performance measures. In *International Conference on Machine Learning (ICML)*, 2005.

[KK07] Richard M Karp and Robert Kleinberg. Noisy binary search and its applications. In *ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2007.

[KNMS08] Robert Kleinberg, Alexandru Niculescu-Mizil, and Yogeshwer Sharma. Regret bounds for sleeping experts and bandits. In *Conference on Learning Theory (COLT)*, 2008.

[Lan08] John Langford. How do we get weak action dependence for learning with partial observations? http://hunch.net/?p=421, September 2008. Blog entry at *Machine Learning (Theory)*.

[LR85] T. L. Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4–22, 1985.

[LS07] Phil Long and Rocco Servedio. Boosting the area under the roc curve. In *Proceedings of Neural Information Processing Systems (NIPS)*, 2007.

[LZ07] John Langford and Tong Zhang. The epoch-greedy algorithm for contextual multi-armed bandits. In *Proceedings of Neural Information Processing Systems (NIPS)*, 2007.

[MR95] Rajeev Motwani and Prabhakar Raghavan. *Randomized Algorithms*. Cambridge University Press, 1995.

[MT04] Shie Mannor and John N. Tsitsiklis. The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research (JMLR)*, 5:623–648, 2004.

[PACJ07] Sandeep Pandey, Deepak Agarwal, Deepayan Chakrabarti, and Vanja Josifovski. Bandits for taxonomies: A model-based approach. In *SIAM Conference on Data Mining (SDM)*, 2007.

[RKJ08] Filip Radlinski, Madhu Kurup, and Thorsten Joachims. How does clickthrough data reflect retrieval quality? In *ACM Conference on Information and Knowledge Management (CIKM)*, 2008.

[Rob52] Herbert Robbins. Some Aspects of the Sequential Design of Experiments. *Bull. Amer. Math. Soc.*, 58:527–535, 1952.

[YJ09] Yisong Yue and Thorsten Joachims. Interactively optimizing information retrieval systems as a dueling bandits problem. In *International Conference on Machine Learning (ICML)*, 2009.