The K-armed Dueling Bandits Problem

Yisong Yue*, Josef Broder**, Robert Kleinberg*, Thorsten Joachims*

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 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

*Department of Computer Science, Cornell University, Ithaca, NY, 14853
**Center for Applied Mathematics, Cornell University, Ithaca, NY, 14853
Email addresses: yyue@cs.cornell.edu (Yisong Yue), jbroder@cam.cornell.edu (Josef Broder), rdk@cs.cornell.edu (Robert Kleinberg), tj@cs.cornell.edu (Thorsten Joachims)

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]. This paper extends results originally published in [YBKJ09] with empirical evaluations as well as a more thorough theoretical analysis.

A canonical application example of the Dueling Bandits Problem 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 to maximize user satisfaction. 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. We empirically evaluate the behavior of both algorithms in Section 7.

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 both the stochastic case [ACBF02] and the non-stochastic case [AB09]. 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 proposed 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 6.

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 cannot be improved. Our dueling bandits problem is a special case of the partial monitoring problem. In particular, 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 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 6.

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 and convex 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 gen-

erated using pairs of inputs with labels from different ordinal classes (e.g., [AM08, BBB+07, FISS03, HGO99, Joa05, LS07]). In the case where there are exactly two label classes, this becomes the so-called bipartite ranking problem [AM08, BBB+07], which is a more general version of learning to optimize ROC-Area [HGO99, Joa05, LS07].

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 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^*,

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

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. Building from this perspective, we can also define **generalized regret**,

$$\bar{R}_T = \sum_{t=1}^{T} r_t(b_1^{(t)}, b_2^{(t)}), \tag{3}$$

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

where

$$r_t(b_1^{(t)}, b_2^{(t)}) \in \left[\min\{\epsilon(b^*, b_1^{(t)}), \epsilon(b^*, b_2^{(t)})\},\ \max\{\epsilon(b^*, b_1^{(t)}), \epsilon(b^*, b_2^{(t)})\}\right].$$

At each time step t, $r_t(b_1^{(t)}, b_2^{(t)})$ is the (potentially non-deterministic) incurred regret of comparing $b_1^{(t)}$ and $b_2^{(t)}$ and is assumed to be bounded between the two individual regret values. Note that both strong regret and weak regret are special cases where $r_t(b_1^{(t)}, b_2^{(t)}) = (\epsilon(b^*, b_1^{(t)}) + \epsilon(b^*, b_2^{(t)}))/2$ and $r_t(b_1^{(t)}, b_2^{(t)}) = \min\{\epsilon(b^*, b_1^{(t)}), \epsilon(b^*, b_2^{(t)})\}$, respectively. We will present algorithms which achieve identical regret bounds for all three 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

$$\epsilon_{i,k} \geq \max\{\epsilon_{i,j}, \epsilon_{j,k}\}. \tag{4}$$

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

$$\epsilon_{i,k} \leq \epsilon_{i,j} + \epsilon_{j,k}. \tag{5}$$

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. We will describe and justify a more general family of probabilistic models in Appendix A.

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

Algorithm 1 Explore Then Exploit Solution

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

4. Algorithm and 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[3].

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 $1/T$, in which case we assume maximal regret $\mathcal{O}(T)$. We can thus bound the expected regret by

$$\begin{aligned} \mathbf{E}[R_T] &\leq \left(1 - \frac{1}{T}\right) \mathbf{E}\left[R_T^{IF}\right] + \frac{1}{T}\mathcal{O}(T) \\ &= \mathcal{O}\left(\mathbf{E}\left[R_T^{IF}\right] + 1\right) \end{aligned} \tag{6}$$

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

[3] In practice, we can terminate EXPLORE after it has run for T time steps, in which case the incurred regret is strictly less than running EXPLORE to completion.

(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 without loss of generality 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 probability of making a mistake is at most $1/T$. We finally bound the matches played by IF1 and IF2 to arrive at our final regret bounds.

4.1. Confidence Intervals

In a match between b_i and b_j, Interleaved Filter maintains a number

$$\hat{P}_{i,j} = \frac{\#\ b_i\ wins}{\#\ comparisons}, \tag{7}$$

which is the empirical estimate of $P(b_i \succ b_j)$ after t comparisons[4]. For ease of notation, we drop the subscripts (b_i, b_j), and use $\hat{P}_t$, which emphasizes the

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

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 $P(\hat{b} > b)$ according to (7)
$\forall b \in W$, maintain $1-\delta$ confidence interval $\hat{C}_{\hat{b},b}$ of $\hat{P}_{\hat{b},b}$ according to (8), (9)
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**
    $W \leftarrow W \setminus \{b\}$ *// $\hat{b}$ declared winner against 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**
    $\hat{b} \leftarrow b'$, $W \leftarrow W \setminus \{b'\}$ *// b' declared winner against $\hat{b}$ (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 $(\hat{b}, \hat{T})$

dependence on the number of comparisons. IF also maintains a confidence interval

$$\hat{C}_t = (\hat{P}_t - c_t, \hat{P}_t + c_t), \tag{8}$$

where

$$c_t = \sqrt{4\log(1/\delta)/t}. \tag{9}$$

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 probability that the inferior bandit is declared the winner at some time $t \leq T$ is at most δ.

Proof. First we argue that the probability of the inferior bandit being declared the winner is at most δ. Note that by the stopping condition of the match, if

we mistakenly declare the inferior bandit the winner at time t, then we must have $1/2 + \epsilon_{i,j} \notin \hat{C}_t$ (note that $\epsilon_{i,j}$ can be either positive or negative). By the definition of $\hat{C}_t$ and the fact that $\mathbf{E}[\hat{P}_t] = 1/2 + \epsilon_{i,j}$, we have $P(1/2 + \epsilon_{i,j} \notin \hat{C}_t) = P(|\hat{P}_t - \mathbf{E}[\hat{P}_t]| \geq c_t)$. It follows from Hoeffding's inequality [Hoe63] that the probability of making a mistake at time t is bounded above by

$$P(|\hat{P}_t - \mathbf{E}[\hat{P}_t]| \geq c_t) \leq 2\exp(-2tc_t^2) = 2\exp(-8\log(1/\delta)) = 2\delta^8 = \frac{2}{T^8K^{16}}.$$

Now an application of the union bound shows that the probability of making a mistake at any time $t \leq T$ is bounded above by

$$P\left(\bigcup_{t=1}^{T}\{1/2 + \epsilon_{i,j} \notin \hat{C}_t\}\right) \leq \frac{2T}{T^8K^{16}} \leq \frac{1}{TK^2} = \delta,$$

provided that $K \geq 2$, which is the desired result.

We now show that the number of comparisons n in a match between b_i and b_j is $\mathcal{O}(\log(TK)/\epsilon_{i,j}^2)$ with high probability. Specifically, we will show that for any $d \geq 1$, there exists an m depending only on d such that

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

for all K sufficiently large. By the stopping condition of the match, if at any time t we have $\hat{P}_t - c_t > 1/2$, then the match terminates. It follows that for any time t, if $n > t$, then $\hat{P}_t - c_t \leq 1/2$, and so

$$P(n > t) \leq P(\hat{P}_t - c_t \leq 1/2).$$

To bound this probability, assume without loss of generality that $\epsilon_{i,j} > 0$, and note that since $\mathbf{E}[\hat{P}_t] = 1/2 + \epsilon_{i,j}$, we have

$$P(\hat{P}_t - c_t \leq 1/2) = P(\hat{P}_t - 1/2 - \epsilon_{i,j} \leq c_t - \epsilon_{i,j}) = P(\mathbf{E}[\hat{P}_t] - \hat{P}_t \geq \epsilon_{i,j} - c_t).$$

For any $m \geq 8$ and $t \geq \lceil 2m\log(TK^2)/\epsilon_{i,j}^2 \rceil$, we have $c_t \leq \epsilon_{i,j}/2$, and so applying Hoeffding's inequality for this m and t shows

$$P(\mathbf{E}[\hat{P}_t] - \hat{P}_t \geq \epsilon_{i,j} - c_t) \leq P(|\hat{P}_t - \mathbf{E}[\hat{P}_t]| \geq \epsilon_{i,j}/2) \leq 2\exp(-t\epsilon_{i,j}^2/2).$$

Since $t \geq 2m\log(TK^2)/\epsilon_{i,j}^2$ by assumption, we have $t\epsilon_{i,j}^2/2 \geq m\log(TK^2)$, and so

$$2\exp(-t\epsilon_{i,j}^2/2) \leq 2\exp(-m\log(TK^2)) = \frac{2}{T^mK^{2m}} \leq K^{-m}$$

for $K \geq 2$, which proves the claim.

□

4.2. Regret per Match

We now bound the accumulated regret of each match. We first bound strong and weak regret, and then extend the result to generalized regret.

Lemma 2. *Assuming b_1 has not been removed and $T \geq K$, then with high probability the accumulated weak regret and also 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

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

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

$$\begin{aligned}
\epsilon_{1,j}\mathcal{O}\left(\frac{1}{\epsilon_{1,j}^2}\log(TK)\right) &= \mathcal{O}\left(\frac{1}{\epsilon_{1,j}}\log(TK)\right) \\
&\leq \mathcal{O}\left(\frac{1}{\epsilon_{1,2}}\log(TK)\right) \\
&= \mathcal{O}\left(\frac{1}{\epsilon_{1,2}}\log T\right) \qquad (10)
\end{aligned}$$

where (10) holds since $\log(TK) \leq \log(T^2) = 2\log T$. 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 $\epsilon_{1,j} + \epsilon_{1,i} \leq 2\epsilon_{1,j}$, and the accumulated strong regret of the match is bounded by

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

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

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

and the accumulated strong regret is bounded by

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

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

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

Like in the analysis for weak regret (10), we finally note that

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

□

Lemma 3. *Assuming b_1 has not been removed and $T \geq K$, then with high probability the accumulated generalized 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. At each time step t that b_i is compared to b_j, the accumulated generalized regret for that comparison is $r(b_i, b_j) \in [\min\{\epsilon_{1,i}, \epsilon_{1,j}\}, \max\{\epsilon_{1,i}, \epsilon_{1,j}\}]$. Let n denote the number of comparisons made in the match. Then the accumulated generalized regret can be bounded by

$$n\max\{\epsilon_{1,i}, \epsilon_{1,j}\} \leq n\left(\epsilon_{1,i} + \epsilon_{1,j}\right) = \mathcal{O}\left(\frac{1}{\epsilon_{1,2}}\log T\right),$$

where the last equality is the regret bound for strong regret derived in Lemma 2. □

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 ($T \geq K$), and IF1 incurs expected generalized regret (and thus also weak and strong regret) bounded by*

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

The theorem will follow from combining Lemma 3, (6), and Lemmas 4 and 6 to follow. We begin by analyzing the probability of IF1 making a mistake.

Lemma 4. *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 (6), and we are interested here in bounding R_T^{IF1}. We can model the sequence of candidate bandits using the following random walk model.

Definition 1. ***(Random Walk Model)*** *Define a random walk graph with K nodes labeled $b_1, \ldots, b_K$ (these will correspond to the similarly named bandits). Each node b_j ($j > 1$) transitions to b_i for $j > i \geq 1$ with probability $1/(j-1)$, or in other words b_j transitions to $b_1, \ldots, b_{j-1}$ with uniform probability. The final node b_1 is an absorbing node.*

A path in the Random Walk Model corresponds to a sequence of candidate bandits taken by IF (both IF1 and IF2) in an instance of the Dueling Bandits problem where $\epsilon_{1j} = \epsilon_{2j} = \ldots = \epsilon_{j-1,j}$ for all $j > 1$ (and no mistakes are made). Thus, the path length of the random walk is exactly to the number of rounds in IF.

Proposition 1. *Either IF makes a mistake, or else the number of rounds in the execution of IF is stochastically dominated by the path length of a random walk in the Random Walk Model.*

Proposition 1 follows directly from Lemma 14 in Appendix B. This allows us to concentrate our analysis on the (simpler) upper bound setting of the Random Walk Model. We will prove that the random walk in the Random Walk Model 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. We first analyze the marginal probability of each $P(X_i = 1)$, and also show that $X_1, \ldots, X_{K-1}$ are mutually independent.

Lemma 5. *Let X_i be as defined above with $1 \leq i < K$. Then*

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

and furthermore, for all $W \subseteq \{X_1, \ldots, X_{K-1}\}$*, we can write* $P(W) \equiv P(\bigwedge_{i \in W} X_i)$ *as*

$$P(W) = \prod_{X_i \in W} P(X_i), \tag{11}$$

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

Proof. We can rewrite (11) as

$$P(W) = \prod_{X_i \in W} P(X_i | W_i),$$

where $W_i = \{X_j \in W | j > i\}$.

We first consider $W = \{X_1, \ldots, X_{K-1}\}$. For the factor on X_i, denote with j the smallest index in W_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 W. We use $\sum_{W^c}$ to indicate summing over the joint states of all X_i variables not in W. We can write $P(W)$ as

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

This proves mutual independence (11).

□

We can express the number of steps taken by a random walk from b_K to b_1 in the Random Walk Model as

$$S_K = 1 + \sum_{i=1}^{K-1} X_i. \tag{12}$$

Lemma 5 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 6. *Assuming IF1 is mistake-free, then it runs for $\mathcal{O}(\log K)$ rounds with high probability.*

Proof. Due to Proposition 1, it suffices to analyze the distribution of path lengths in the Random Walk Model. It thus suffices to show that for any d sufficiently large, there exists a m depending only on d such that

$$\forall K \geq 1 : \qquad P(S_K > m \log K) \leq \frac{1}{K^d}, \tag{13}$$

for S_K as defined in (12). From Lemma 5, we know that the random variables $X_1, \ldots, X_{K-1}$ in S_K are mutually independent. Then using the Chernoff bound [MR95], we know that for any $m > 1$,

$$\begin{aligned} P(S_K > m(1 + H_{K-1})) \quad &\leq \left(\tfrac{e^{m-1}}{m^m}\right)^{1+H_{K-1}} \\ &\leq \left(\tfrac{e^{m-1}}{m^m}\right)^{1+\log K} \\ &= (eK)^{m-1-m\log m} \end{aligned} \tag{14}$$

(14) is true since

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

for all $K \geq 1$. We require this bound to be at most $1/K^d$, or

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

The above inequality is satisfied by $m \geq d$ for $d \geq e$. The Chernoff bound applies for all $K \geq 0$. So for any $d \geq e$, we can choose $m = d$ to satisfy (13). □

Corollary 1. *Assuming IF1 is mistake-free, then it plays $\mathcal{O}(K \log K)$ matches with high probability.*

Proof. The result immediately follows from Lemma 6 by noting that IF1 plays at most $\mathcal{O}(K)$ matches in each round. □

Algorithm 3 Interleaved Filter 2 (IF2)

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 $P(\hat{b} > b)$ according to (7)
$\forall b \in W$, maintain $1-\delta$ confidence interval $\hat{C}_{\hat{b},b}$ of $\hat{P}_{\hat{b},b}$ according to (8), (9)
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**
    $W \leftarrow W \setminus \{b\}$ *//$\hat{b}$ declared winner against 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
    $\hat{b} \leftarrow b'$, $W \leftarrow W \setminus \{b'\}$ *//b' declared winner against $\hat{b}$ (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 $\mathcal{B} = \{b_1, \ldots, b_K\}$, time horizon T ($T \geq K$), and IF2 incurs expected generalized regret (and thus also weak and strong regret) 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 3, (6), and Lemmas 8 and 9 to follow. IF2 improves upon IF1 by removing all empirically inferior bandits whenever the incumbant is defeated, which we call pruning. We begin by analyzing the pruning technique. The following lemma could be informally summarized by saying that when IF2 produces a new incumbent b' and then eliminates a bandit b in the subsequent pruning step, we can conclude that b' is superior to b with $1-(\delta T)$ confidence.

Lemma 7. *For all triples of bandits $b, b', \hat{b}$ such that $b \succ b'$, the probability that IF2 eliminates b in a pruning step in which b' wins a match against the incumbent bandit $\hat{b}$ (i.e. $\hat{P}_{\hat{b},b'} < 1/2$) while b is found to be empirically inferior to $\hat{b}$ (i.e. $\hat{P}_{\hat{b},b} > 1/2$) is at most δ.*

Proof. Let $X_1, X_2, \ldots$ denote an infinite sequence of i.i.d. Bernoulli random variables with $\mathbf{E}[X_i] = P(\hat{b} \succ b')$, and let $Y_1, Y_2, \ldots$ denote an infinite sequence of i.i.d. Bernoulli random variables with $\mathbf{E}[Y_i] = P(\hat{b} \succ b)$. We couple the outcomes of the comparisons performed by the algorithm to the sequences (X_i), (Y_i) in the obvious way: X_i (resp. Y_i) represents the outcome of the i^{th} comparison between $\hat{b}$ and b' (resp. $\hat{b}$ and b) if the algorithm performs at least i comparisons of that pair of bandits; otherwise X_i (resp. Y_i) does not correspond to any comparison observed by the algorithm.

If b is eliminated by IF2 in a pruning step at the end of a match consisting of n comparisons between b' and the incumbent $\hat{b}$, then $X_1, \ldots, X_n$ represent the outcomes of the n matches between $\hat{b}$ and b' in that round, and $Y_1, \ldots, Y_n$ represent the outcomes of the n matches between $\hat{b}$ and b in that round. From the definition of confidence intervals in IF2 we know that $X_1 + \cdots + X_n < n/2 - \sqrt{4n \log(1/\delta)}$, whereas the definition of the pruning step implies that $Y_1 + \cdots + Y_n > n/2$. Thus, if we define $Z_i = Y_i - X_i$ for $i = 1, 2, \ldots$, then we have

$$Z_1 + \cdots + Z_n > \sqrt{4n \log(1/\delta)}. \tag{15}$$

To complete the proof of the lemma, we will show the probability that there exists an n satisfying (15) is at most δT.

The random variables $(Z_i)_{i=1}^{\infty}$ are i.i.d. and satisfy $|Z_i| \leq 1$. Furthermore, our assumption that $b \succ b'$ together with strong stochastic transitivity implies that

$$\mathbf{E}[Z_i] = P(\hat{b} \succ b) - P(\hat{b} \succ b') \leq 0.$$

By Hoeffding's inequality, for every n the probability that $\sum_{i=1}^{n} Z_i$ exceeds $\sqrt{4n \log(1/\delta)}$ is at most $\exp(-8n \log(1/\delta)/(4n)) = \delta^2$. Taking the union bound over $n = 1, 2, \ldots, T$, we find that the probability that there exists an n satisfying (15) is at most $\delta^2 T \leq \delta$, as claimed. □

Lemma 8. *The probability that IF2 makes a mistake resulting in the eliminiation of bandit b_1 is at most $1/T$.*

Proof. By Lemma 1, for every i the probability that b_1 is eliminated in a match against b_i is at most δ. A union bound over all i implies that the probability of b_1 being eliminated by directly losing a match to some other bandit is at most $\delta(K-1)$. On the other hand, by Lemma 1, for all i, j the probability that b_1 is eliminated in a pruning step resulting from a match in which b_i defeats b_j is at most δ. A union bound over all i, j implies that the probability of b_1 being eliminated in a pruning step is at most $\delta(K-1)^2$. Summing these two bounds, the probability that IF2 makes a mistake resulting in the elimination of b_1 is at most $\delta[(K-1) + (K-1)^2] < \delta K^2 = 1/T$. □

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

Proof. Let B_j denote a random variable counting the number of matches played by b_j when it is *not* the incumbant (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 incumbant 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 incumbant was superior to b_j). We can thus bound the expected number of matches played via

$$\sum_{j=1}^{K-1}\mathbf{E}[B_j] = \sum_{j=1}^{K-1}\mathbf{E}[A_j] + \mathbf{E}[G_j]. \tag{16}$$

By Lemma 5 and leveraging the Random Walk Model defined in Section 4.3, we can write $\mathbf{E}[A_j]$ as

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

where H_i is the harmonic sum.

We now analyze $\mathbf{E}[G_j]$. We assume the worst case that b_j does not lose a match (with $1-\delta$ confidence) to any superior incumbant 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 incumbant 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 incumbant, 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

$$\mathbf{E}[G_j] \leq \mathbf{E}[G_{j,1}] \leq 1 + P(\mathcal{E}_{j,1}^c)\mathbf{E}[G_{j,2}]. \tag{17}$$

We know that $P(\mathcal{E}_{j,t}^c) \leq 1/2$ for all $j \neq 1$ and t. From Lemma 6, we know that $\mathbf{E}[G_{j,t}] \leq \mathcal{O}(K\log K)$ and is thus finite. Hence, we can bound (17) by the infinite geometric series $1 + 1/2 + 1/4 + \ldots = 2$.

We can thus write (16) as

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

□

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 3. *Any algorithm ϕ for the dueling bandits problem satisfies*

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

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

Here is a heuristic explanation of why we might suspect the theorem to be true. Rather than consider the general problem of identifying the best of K bandits, suppose we are given a bandit b, and asked to determine with probability at least $1 - 1/T$ whether b is the best bandit. (Intuitively, the regret incurred by the optimal algorithm for this decision problem should be a lower bound on the regret incurred by the optimal algorithm for the general problem). We have seen that, given two bandits b_i and b_j with $P(b_i > b_j) = 1/2 + \epsilon$, we can identify the better bandit with probability at least $1 - 1/T$ after $O(\log T/\epsilon^2)$ comparisons. If this is in fact the minimum number of comparisons required, then we would suspect that any algorithm for the above decision problem that is uniformly good over all problem instances must perform $\Omega(\log T/\epsilon^2)$ comparisons involving each inferior bandit. We will see in Lemma 10 that this is in fact the case, and we begin by constructing the appropriate problem instance.

Fix $\epsilon > 0$ and define the following family of problem instances. In instance j, let b_j be the best bandit, and order the remaining bandits by their indices. That is, in instance j, we have $b_j \succ b_k$ for all $k \neq j$, and for $i, k \neq j$, we have $b_i \succ b_k$ whenever $i < k$. Given this ordering, define the winning probabilities

by $P(b_i > b_k) = 1/2 + \epsilon$ whenever $b_i \succ b_k$. Note that this construction yields a valid problem instance, i.e. one that satisfies (4), (5).

Let q_j be the distribution on T-step histories induced by a given algorithm ϕ under instance j, and let $n_{j,T}$ be the number of comparisons involving bandit b_j scheduled by ϕ up to time T. Using these instances, we prove Lemma 10, from which Theorem 3 follows.

Lemma 10. *Let ϕ be an algorithm for the dueling bandits problem such that*

$$R_T^{\phi} = o(T^a) \tag{18}$$

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

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

Lemma 10 formalizes the intuition given above, in that any algorithm whose regret is $o(T^a)$ over all problem instances must make $\Omega(\log T/\epsilon^2)$ comparisons involving each inferior bandit, in expectation. The proof is motivated by Lemma 5 of [KNMS08].

Proof. Fix an algorithm ϕ satisfying assumption (18), and fix $0 < a < 1/2$. Define the event $\mathcal{E}_j = \{n_{j,T} < \log(T)/\epsilon^2\}$, and let $J = \{j : q_1(\mathcal{E}_j) < 1/3\}$. For each $j \in J$, we have by Markov's inequality that

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

so it remains to show that $\mathbf{E}_{q_1}[n_{j,T}] = \Omega(\log T/\epsilon^2)$ for each $j \notin J$. For any j, we know that under q_j, the algorithm ϕ incurs regret ϵ for every comparison involving a bandit $b \neq b_j$. This fact together with the assumption (18) on ϕ implies that $\mathbf{E}_{q_j}[T - n_{j,T}] = o(T^a)$. Using this fact and Markov's inequality, we have

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

and so choosing T sufficiently large shows that $q_j(\mathcal{E}_j) < 1/3$ for each j (and in particular, that $1 \in J$ by construction). Now by Lemma 6.3 of [KK07], we have that for any event $\mathcal{E}$ and distributions p, q with $p(\mathcal{E}) \geq 1/3$ and $q(\mathcal{E}) < 1/3$,

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

For each $j \notin J$, we may apply this lemma with q_1, q_j, and the event $\mathcal{E}_j$, to show

$$\begin{aligned} KL(q_1; q_j) &\geq \frac{1}{3}\ln\left(\frac{1}{3o(T^{a-1})}\right) - \frac{1}{e} \\ &= \Omega(\log T). \end{aligned} \tag{19}$$

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

$$\begin{aligned} KL(q_1; q_j) &\leq \mathbf{E}_{q_1}[n_{j,T}]\, KL(1/2+\epsilon; 1/2-\epsilon) \\ &\leq 16\epsilon^2 \mathbf{E}_{q_1}[n_{j,T}], \end{aligned} \tag{20}$$

where we use the shorthand $KL(1/2+\epsilon; 1/2-\epsilon)$ to denote the KL-divergence between two Bernoulli distributions with parameters $1/2+\epsilon$ and $1/2-\epsilon$, respectively. The first inequality follows from the fact that the distribution on the outcome of a comparison will differ under distributions q_1 and q_j only if the comparison involves bandit b_j, and the second inequality follows from a standard result on the KL divergence between two Bernoulli distributions. Combining (19) and (20) shows that $\mathbf{E}_{q_1}[n_{j,T}] = \Omega(\log T/\epsilon^2)$ for each $j \notin J$, which proves the lemma. □

Proof of Theorem 3. Let ϕ be any algorithm for the dueling bandits problem. If ϕ does not satisfy the hypothesis of Lemma 10, 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 10 that

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

□

6. 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. $P(b^* > b_{i,j,k}) \approx 1$, but $P(b_i > b_j) = 1/2+\gamma$ for $\gamma \ll 1$). 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.

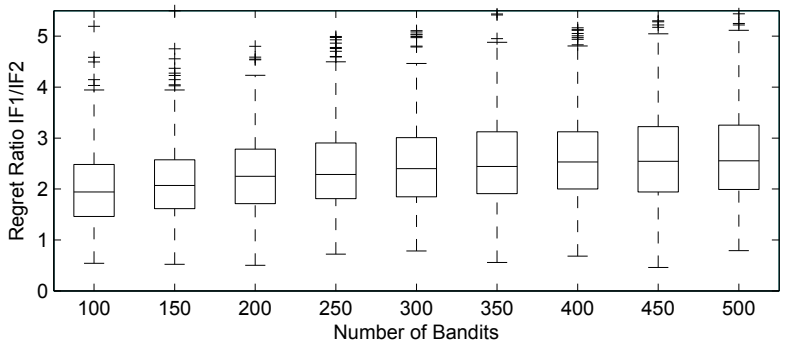

Figure 1: Comparing regret ratio between IF1 and IF2 in worst-case simulations.

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.

7. Experiments

7.1. Synthetic Simulations

We performed numerical simulations on two synthetic problem instances. The first set of simulations used the worst-case instance from the lower bound proof of Theorem 3. In this instance, $P(b_i > b_j) = 1/2 + \epsilon$ whenever $i < j$. For the experiment, we fixed $\epsilon = 0.1$ and a time horizon $T = 10^7$. We varied K from 100 to 500 in increments of 50, and for each value of K, we performed 500 simulations of both IF1 and IF2. In Figure 1, we plot the ratio of the regret incurred by IF1 and IF2 (which we henceforth also call the regret ratio).

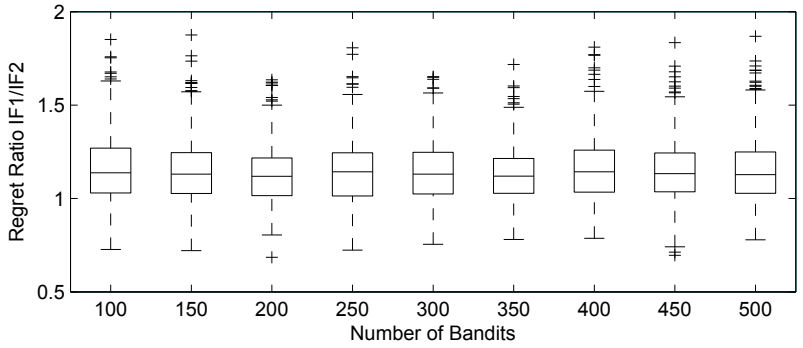

Figure 2: Comparing regret ratio between IF1 and IF2 in random case simulations.

For the second set of simulations, we generated random problem instances according to a Bradley-Terry model with uniformly random weights. For normalization purposes, we then modified each problem instance to ensure that the best bandit had a winning probability of at least $1/2 + \epsilon$ against all other bandits. The details of the procedure are as follows. For each value of K, we generated $K - 1$ random numbers $w_2, \ldots, w_K$ sampled independently from the uniform distribution on $(0, 1)$. To define w_1, we found the largest weight $w_{\max} = \max\{w_2, \ldots, w_K\}$, and defined $w_1 = w_{\max}(1 + 2\epsilon)/(1 - 2\epsilon)$. We then defined $P(b_i > b_j) = w_i/(w_i + w_j)$, so that for all $i \neq 1$,

$$P(b_1 > b_i) = \frac{w_1}{w_1 + w_i} \geq \frac{w_1}{w_1 + w_{\max}} = \frac{1}{2} + \epsilon.$$

Note that this is the Bradley-Terry model discussed in Section 3.1, which satisfies the assumptions of that section. We fixed $\epsilon = 0.1$ and $T = 10^7$, and performed 500 simulations of IF1 and IF2 on each of the randomly generated instances. We plot the regret ratio of IF1 and IF2 in Figure 2.

For the worst-case simulations, we see that IF2 outperforms IF1, and that the median of the regret ratio increases logarithmically with K. For the random-case simulations, we see that IF2 outperforms IF1, but the regret ratio does not increase with K as in the worst-case simulations. Intuitively, IF1 and IF2 incur a large amount of regret during matches in which $P(b_i > b_j)$ is close to $1/2$. In the worst-case problem instance, this is guaranteed to be true for every match, by construction. Consequently, each pruning step performed by IF2 reduces the total regret incurred by a significant amount, by eliminating a high-cost match that would otherwise be played. In contrast to the worst-case instances, we expect that in the random-case, many matches will have $P(b_i > b_j)$ far from $1/2$, and thus will contribute little to the total regret. A pruning step that eliminates such a match will have little effect on the total regret, and so we should expect the regret of IF1 and IF2 to be more similar in the random-case than in the worst-case, as observed in our simulations.

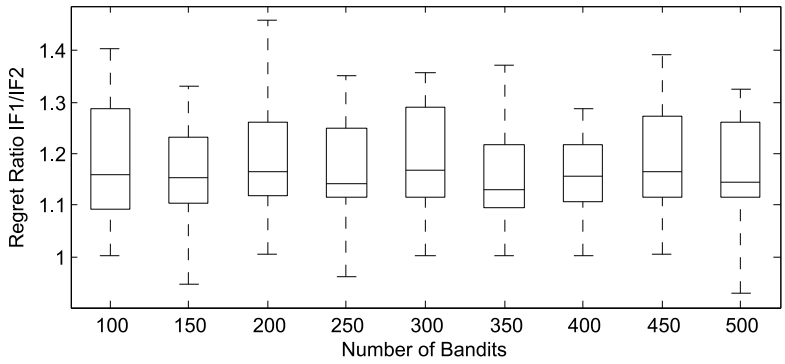

Figure 3: Comparing regret ratio between IF1 and IF2 in web search simulations.

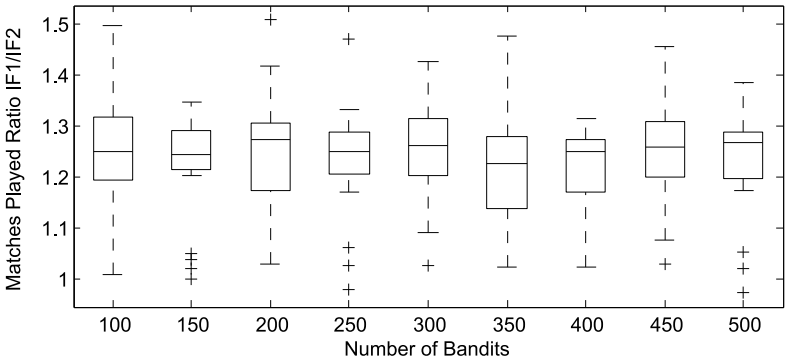

Figure 4: Comparing matches played ratio between IF1 and IF2 in web search simulations.

7.2. Web Search Simulations

For a more realistic simulation, we leveraged a real Web Search dataset (courtesy of Chris Burges at Microsoft Research). The idea is to simulate users issuing queries by sampling from queries in the dataset. For each query, the competing retrieval functions will produce rankings, after which the "user" will randomly prefer one ranking over the other. User preferences are modeled probabilistically using the logistic transfer function and NDCG@10, which is a measure used for evaluating the quality of rankings in information retrieval tasks (see [DSB09]).

We varied the number of bandits (retrieval functions) K from 100 to 500 in increments of 50. For each experimental setting, we randomly selected K retrieval functions from a pool of 1000 retrieval functions gathered from similar experiments performed for the continuous dueling bandits problem [YJ09]). For each value of K, we used 25 experimental settings with 25 trials per setting. We fixed $T = 10^7$ for all settings, since our primary goal in this experiment is to compare the performance of IF1 and IF2. We used strong regret (1) to measure

performance.

Figure 3 shows a box plot of the regret ratio for IF1 and IF2. Since different collections of retrieval functions yield different performances (due to differences in the distinguishability between the bandits), it is more informative to compare the ratio of regret on the same initial conditions[5]. We can see that IF2 consistently outperforms IF1, however the performance ratio does not scale as $log(K)$ as implied by our worst case bounds.

Intuitively, as also discussed for the synthetic experiments, there are two conditions that must be satisfied for IF2 to improve by a logarithmic factor over IF1. First, a logarithmic number of rounds must be played (i.e., we must consider a logarithmic number of candidate bandits). Second, within each round, most of the bandits must not be confidently eliminated from consideration (so they can be eliminated via the pruning procedure in IF2). Satisfying both of these conditions would imply IF1 playing a logarithmic factor more matches than IF2. In the web search dataset, we observe neither condition being strongly satisfied. In all settings, only a small number of rounds are played (typically between 2 and 4) for all values of K (which admittedly only ranges up to 500 in our experiments). Futhermore, in many rounds, a substantial fraction of the bandits are confidently eliminated from consideration before the conclusion of the round. This is summarized in Figure 4, which shows a box plot of the ratio of matches played between IF1 and IF2. Nonetheless, we can see that IF2 can offer real practical improvements over IF1, although the difference in performance is perhaps not as dramatic as suggested by the worst case analysis.

8. 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 are 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 Awards IIS-0812091 and IIS-0905467. The first author is also supported by a Microsoft Research Graduate Fellowship and a

[5]Both IF1 and IF2 start with the same initial incumbant bandit and the same (randomly selected) permutation ordering over the remaining bandits to use for when interleaving matches in round robin fashion.

Yahoo! Key Scientific Challenges Award. The third author is supported by NSF Awards CCF-0643934 and AF-0910940, an Alfred P. Sloan Foundation Fellowship, and a Microsoft Research New Faculty Fellowship.

[AB09] Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *Conference on Learning Theory (COLT)*, 2009.

[ACBF02] Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed 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.

[DSB09] P. Donmez, K. Svore, and C. Burges. On the Local Optimality of LambdaRank. In *Proceedings of ACM Conference on Research and Development in Information Retrieval (SIGIR)*, 2009.

[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.

[YBKJ09] Yisong Yue, Josef Broder, Robert Kleinberg, and Thorsten Joachims. The k-armed dueling bandits problem. In *Conference on Learning Theory (COLT)*, 2009.

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

Appendix A. Satisfying Modeling Assumptions

The following lemma describes a general family of probabilistic comparison models and proves that strong stochastic transitivity and stochastic triangle inequality are both satisfied by this family of models. Note that both the logistic and Gaussian models described in Section 3.1 are contained within this family of models.

Lemma 11. *Let each bandit $b_i \in \{b_1 \dots b_K\}$ be associated with a distinct real value μ_i such that outcomes from comparing two bandits are determined by*

$$P(b_i > b_j) = \sigma(\mu_i - \mu_j),$$

for some transfer function σ. Let σ satisfy the following properties:

- *σ is monotonically increasing*
- $\sigma(-\infty) = 0$
- $\sigma(\infty) = 1$
- $\sigma(x) = 1 - \sigma(-x)$ *(rotation symmetric)*
- *$\sigma(x)$ has a single inflection point at $\sigma(0) = 1/2$*

Then these probabilistic comparisons satisfy strong stochastic transitivity and stochastic triangle inequality.

Proof. We begin by noting that that the properties assumed about σ essentialy indicates that σ behaves like a symmetric cumulative distribution function with a single point inflection point at $\sigma(0) = 1/2$ (i.e., σ is an "S-shaped" curve).

For any triplet of bandits $b_i \succ b_j \succ b_k$, we know that $\mu_i > \mu_j > \mu_k$. To show strong stochastic transitivity, we note that σ is monotonically increasing, Thus we know that $\sigma(\mu_i - \mu_k) \geq \sigma(\mu_i - \mu_j)$ and $\sigma(\mu_i - \mu_k) \geq \sigma(\mu_j - \mu_k)$, which implies that

$$\epsilon_{i,k} = \sigma(\mu_i - \mu_k) - \frac{1}{2} \geq \max\left\{\sigma(\mu_i - \mu_j) - \frac{1}{2}, \sigma(\mu_j - \mu_k) - \frac{1}{2}\right\} = \max\left\{\epsilon_{i,j}, \epsilon_{j,k}\right\}.$$

To show stochastic triangle inequality, we first note that $\sigma(x)$ is sub-additive, or concave, for $x > 0$. Define $\alpha \in (0, 1)$ such that $(\mu_i - \mu_j) = \alpha(\mu_i - \mu_k)$ and $(\mu_j - \mu_k) = (1 - \alpha)(\mu_i - \mu_k)$. Then we know from concavity of σ that

$$\alpha\sigma(\mu_i - \mu_k) + (1 - \alpha)\sigma(0) \leq \sigma(\mu_i - \mu_j),$$

and also

$$(1 - \alpha)\sigma(\mu_i - \mu_k) + \alpha\sigma(0) \leq \sigma(\mu_j - \mu_k).$$

Add the two inequalities above yields

$$\sigma(\mu_i - \mu_k) + \mu(0) \leq \sigma(\mu_i - \mu_j) + \sigma(\mu_j - \mu_k),$$

and thus

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

□

Appendix B. Analyzing the Random Walk Model

We first describe a family of measure spaces which will be used to analyze the coupling between executions of IF and the Random Walk Model.

Definition 2. *We define a family of measure spaces $\mathcal{M}$ in the following way. Each point in the sample space is a joint realization of the sequences of random variables X_{ij}^{rt} and Z_i^r for every pair of bandits b_i and b_j, and positive integers r and t. We will define a joint distribution over the random variables X_{ij}^{rt} and a conditional distribution over the Z_i^r variables given the X_{ij}^{rt} variables. The random variables and their distributions are explained in greater detail below.*

- *For every pair of bandits b_i, b_j, and positive integer r, there is a sequence of Bernoulli random variables X_{ij}^{rt} (for $t = 1, 2, \ldots$) describing the outcomes of comparisons in a match played by b_i and b_j in round r provided that b_i is the incumbent in that round. In particular $X_{ij}^{rt} = 1$ if b_i wins the t-th comparison between b_j in round r, and $X_{ij}^{rt} = 0$ if b_i loses that comparison. We will also define the following useful notation to denote prior execution histories: $\mathcal{X}_i^r$ is the σ-field generated by the random variables $\{X_{ij}^{qt} : j \neq i, q < r, t = 1, 2, \ldots\}$.*

- *For a fixed i, the random variables X_{ij}^{rt} are all mutually independent as one varies j, r, t, and they have the correct distribution for each pair i, j. (In other words, the probability of b_i beating b_j is $1/2 + \epsilon_{ij}$).*

- *For convenience we also define Y^r, for every positive integer r, to denote the identity of the incumbent in round $r+1$ (i.e., the bandit that wins round r) when running algorithm IF with the comparison outcomes specified by $\{X_{ij}^{rt}\}$. Note that the value (likewise distribution) of Y^r is completely determined by the values (joint distribution) of X_{ij}^{rt}.*

- *For every bandit b_i and positive integer r, there is a random variable Z_i^r taking non-negative integer values, such that the distribution of $Y^r + Z_i^r$, conditioned on $\mathcal{X}_i^r$, is uniform on $1, \ldots, i-1$ at every sample point where $Y^{r-1} \leq i$ and IF does not make a mistake in rounds $1, \ldots, r$. (This will later be used to show that the Random Walk Model stochastically dominates any mistake-free execution of IF.)*

The values of X_{ij}^{rt} completely determine the history of execution of IF[6]. Our independence assumptions ensure that the history of play observed by IF has the correct distribution over histories.

A priori, it is not obvious that measure spaces $\mathcal{M}$ satisfying Definition 2 exist; the constraint on the conditional distribution of $Y^r + Z_i^r$ is non-trivial but we prove below that it is possible to design a measure space that satisfies this constraint, i.e. $\mathcal{M}$ is not empty. We will then show how any measure space in $\mathcal{M}$ defines a stochastic coupling between the number of rounds required in mistake-free executions of IF and the length of random walks in the Random Walk Model $\mathcal{G}$. To begin proving that $\mathcal{M}$ is non-empty, we first prove a constraint on the distribution of the Y^r variables.

Lemma 12. *For any measure space in $\mathcal{M}$, we have*

$$\forall r,\ \forall 1 \leq j < i:\ \sum_{j'=1}^{j} P(Y^r = j' | \mathcal{X}_i^r, N^r) \geq \frac{j}{i-1}, \tag{B.1}$$

where b_i denotes the incumbent bandit chosen by IF for round r, the Y^r and X_{ij}^{rt} variables and the $\mathcal{X}_i^r$ σ-field are defined as in Definition 2, and N^r denotes the event that IF does not make a mistake in round r.

Proof. We will prove the following inequality,

$$\forall t \geq t_{min}, \forall r,\ \forall 1 \leq j < i:\ \sum_{j'=1}^{j} P(Y^r = j' | \mathcal{X}_i^r, N^{rti}) \geq \frac{j}{i-1}, \tag{B.2}$$

[6]Some of the values X_{ij}^{rt} are exposed as IF runs and schedules matches. Other values never get exposed. In particular, if b_i is not the incumbent in round r, then the values X_{ij}^{rt} have no bearing on the history of play observed by IF.

where t_{min} denotes the minimum number of comparisons required for IF to determine a winner, and N^{rti} denotes the event that IF does not make a mistake in round r, that b_i is the incumbent in that round, and that IF makes exactly t comparisons between b_i and each other remaining bandit in round r. Since (B.2) will be shown to apply for all feasible t, i, then (B.1) will also hold.

It suffices to show that

$$\forall 1 \leq j < k < i: \; P(Y^r = j | \mathcal{X}_i^r, N^{rti}) \geq P(Y^r = k | \mathcal{X}_i^r, N^{rti}), \tag{B.3}$$

since then (B.2) follows from iteratively applying the pigeonhole principle (for $j = 1, \ldots, i-1$), and noting that

$$\sum_{j'=1}^{i-1} P(Y^r = j' | \mathcal{X}_i^r, N^{rti}) = 1.$$

Let $U(i, k, r, t | \mathcal{X}_i^r)$ denote the collection of comparison sequences of length t in round r between the incumbent b_i and each other remaining b_j which results in b_k being declared the winner after t comparisons. In other words, an element in $U(i, k, r, t | \mathcal{X}_i^r)$ consists of a realization of each $X_{ij}^{t'r}$ for incumbent b_i, all remaining b_j, and time steps $1 \leq t' \leq t$. It is straightforward to see that

$$P(Y^r = k | \mathcal{X}_i^r, N^{rti}) = P(U(i, k, r, t | \mathcal{X}_i^r) | \mathcal{X}_i^r, N^{rti}).$$

We define a bijection between $U(i, j, r, t | \mathcal{X}_i^r)$ and $U(i, k, r, t | \mathcal{X}_i^r)$ for $j < k$ such that $P(U(i, j, r, t | \mathcal{X}_i^r) | \mathcal{X}_i^r, N_{rt}) \geq P(U(i, k, r, t | \mathcal{X}_i^r) | \mathcal{X}_i^r, N^{rti})$, which directly implies (B.3). Each $u_k \in U(i, k, r, t | \mathcal{X}_i^r, N^{rti})$ is mapped to the corresponding point $u_j \in U(i, j, r, t | \mathcal{X}_i^r, N^{rti})$ that consists of the same sequences of comparisons as u_k, except that the comparison sequences involving b_j and b_k are swapped (implying that b_j is declared the winner).

It remains to show that $P(u_j | \mathcal{X}_i^r, N^{rti}) \geq P(u_k | \mathcal{X}_i^r, N^{rti})$ for all u_j, u_k pairings in the bijection. In the sequences of comparisons defined by u_k, let

$$A = \sum_{t'=1}^{t} X_{ik}^{rt'} \quad \text{and} \quad B = \sum_{t'=1}^{t} X_{ij}^{rt'},$$

where $A > B$. Under the corresponding u_j, the two summations are reversed,

$$B = \sum_{t'=1}^{t} X_{ik}^{rt'} \quad \text{and} \quad A = \sum_{t'=1}^{t} X_{ij}^{rt'},$$

and all other sequences of variables $X_{ii'}^{rt'}$ for $i' \neq j$, $i' \neq k$ remain the same. We also know that $P(X_{ik}^{rt}) \leq P(X_{ij}^{rt})$, since b_k is inferior to b_j. Let $p = P(X_{ij}^{rt})$ and $q = P(X_{ik}^{rt})$. Since all the $X_{ii'}^{rt'}$ variables are mutually independent, we can

write the ratio of the conditional probabilities of u_j and u_k as

$$\begin{aligned}\frac{P(u_j|\mathcal{X}_i^r, N^{rti})}{P(u_k|\mathcal{X}_i^r, N^{rti})} &= \frac{P(\sum_{t=1}^{t'} X_{ij}^{rt} = A)P(\sum_{t=1}^{t'} X_{ik}^{rt} = B)}{P(\sum_{t=1}^{t'} X_{ij}^{rt} = B)P(\sum_{t=1}^{t'} X_{ik}^{rt} = A)} \\ &= \frac{p^A(1-p)^{t'-A}q^B(1-q)^{t'-B}}{p^B(1-p)^{t'-B}q^A(1-q)^{t'-A}} \\ &= \frac{p^{A-B}(1-q)^{A-B}}{q^{A-B}(1-p)^{A-B}} \geq 1\end{aligned}$$

where the first equality follows from noting that all comparisons are independent and canceling out like terms (i.e., the realizations of the comparisons for other $X_{ii'}^{rt}$ where $i' \neq j$ and $i' \neq k$), and the last inequality follows from noting that $A > B$ and $p > q$.

□

Corollary 2. *For the setting described in Lemma 12, we also have*

$$\forall r,\ \forall 1 \leq j < i:\ \sum_{j'=1}^{j} P(Y^r = j'|\mathcal{X}_i^r, N^r) \geq \frac{j}{i'-1},$$

where $i' \geq i$.

Lemma 13. *The family of measure spaces $\mathcal{M}$ defined in Definition 2 is non-empty.*

Proof. We will use the notation for $X_{ij}^{rt}, Y^r, Z_i^r, \mathcal{X}_i^r$ as described in Definition 2. We will show that it is possible to construct a distribution on the non-negative random variables Z_i^r which satisfies the requirements of Definition 2. Since we are conditioning on X_{ij}^{qt} for all $q < r$, then the value of Y^{r-1} is fixed (i.e., we know who the incumbent is in round r). We will construct Z_i^r based on the following two cases.

Case 1: IF does not make a mistake in round r and $Y^{r-1} \leq i$ (meaning the incumbent during round r was b_i). We will use the following flow network to construct the conditional distribution of $Y^r + Z_i^r$ (given $\mathcal{X}_i^r$ and N^r),

- source s and sink t
- vertices $u_1, \ldots, u_{i-1}$
- vertices $v_1, \ldots, v_{i-1}$
- edges from s to each u_j with capacity $P(Y^r = j|\mathcal{X}_i^r, N^r)$
- edges from each u_j to v_k where $k \geq j$ with infinite capacity
- edges from each v_k to t with capacity $1/(i-1)$

Lemma 12 and Corollary 2 imply that the minimum s-t cut of this network has capacity 1, and consequently the maximum s-t flow has value 1. In any maximum flow, each edge (s, u_j) and each edge (v_j, t) (for $1 \leq j \leq i-1$) must be saturated. Given a maximum flow, we can interpret the flow on the edge from u_j to v_k to be the joint conditional probability $P(Y^r = j,\ Z_i^r = k - j \mid \mathcal{X}_i^r, N^r)$, from which we can recover the conditional distribution of Z_i^r given $\mathcal{X}_i^r$ and N^r. The fact that the conditional distribution of $Y^r + Z_i^r$ is uniform on $1, \ldots, i-1$, given $\mathcal{X}_i^r, N^r$, follows from the fact that the flow from v_k to t is exactly $1/(i-1)$ for every k.

Case 2: IF does make a mistake in round r or $Y^{r-1} > i$. Then we set Z_i^r to some arbitrary non-negative integer, e.g., 0.

Thus, we have shown that there exists a feasible probability distribution on the Z_i^r variables which satisfies the requirements of Definition 2, which implies that $\mathcal{M}$ is non-empty. □

Lemma 14. *There exists a stochastic coupling between IF and the Random Walk Model such that the number of rounds in mistake-free executions of IF is stochastically dominated by the length of random walks in the Random Walk Model.*

Proof. We can take any measure space in $\mathcal{M}$ to construct our stochastic coupling, and we know from Lemma 13 that at least one such measure space exists. There is one sample point for every possible joint outcome of the random variables X_{ij}^{rt} and Z_i^r. The execution of IF is determined by the X_{ij}^{rt} variables. Consider any execution of IF that is mistake-free through rounds $1, \ldots, s$. The analogous execution of the Random Walk Model is determined by looking at the sequence of incumbents when one runs a "perturbed" version of IF. The perturbation consists to taking the identity of the incumbent in round $r+1$ (for every $r = 1, \ldots, s$) and modifying it by adding Z_i^r (where b_i is the incumbent of "perturbed" IF in round r), and then executing round $r+1$ using the perturbed incumbent instead of the one that would ordinarily be chosen by IF. Both IF and "perturbed" IF start with the same initial incumbent at the beginning of round 1 chosen uniformly from $1, \ldots, K$.

Let b^r and $\tilde{b}^r$ be the incumbents chosen by IF and the analogous "perturbed" IF, respectively, at round r (note that $b^r = b_{i'}$ where $i' = Y^{r-1}$). Then it suffices to show that any mistake-free execution of IF satisfies $b^r \succeq \tilde{b}^r$ for all $r > 0$. It is straightforward to see that this stochastic coupling holds from the definition of the Y^r and Z_i^r variables in Definition 2, so long the initial condition $b^1 \succeq \tilde{b}^1$ holds (and $b^1 = \tilde{b}^1$ by definition). □