8. Appendix

Here we provide some details that were alluded to in the main body of the paper.

8.1. The Condorcet Assumption

In the K-armed dueling bandit problem, regret is measured with respect to the Condorcet winner. The Condorcet winner differs in a subtle but important way from the *Borda winner* (Urvoy et al., 2013), which is an arm a_b that satisfies $\sum_j p_{bj} \geq \sum_j p_{ij}$, for all $i = 1, \ldots, K$. In other words, when averaged across all other arms, the Borda winner is the arm with the highest probability of winning a given comparison.

In the K-armed dueling bandit problem, the Condorcet winner is sought rather than the Borda winner, for two reasons. First, in many applications, including the ranker evaluation problem addressed in our experiments, the eventual goal is to adapt to the preferences of the users of the system. Given a choice between the Borda and Condorcet winners, those users prefer the latter in a direct comparison, so it is immaterial how these two arms fare against the others. Second, in settings where the Borda winner is more appropriate, no special methods are required: one can simply solve the K-armed bandit algorithm with arms $\{a_1, \ldots, a_K\}$, where pulling a_i means choosing an index $j \in \{1, \ldots, K\}$ randomly and comparing a_i against a_j. Thus, research on the K-armed dueling bandit problem focuses on finding the Condorcet winner, for which special methods are required to avoid mistakenly choosing the Borda winner.

As mentioned in Section 3, IF and BTM assume more than the existence of a Condorcet winner. They also require the comparison probabilities p_{ij} to satisfy certain restrictive and difficult to verify conditions. Specifically, IF and BTM require a *total ordering* $\{a_1, \ldots, a_K\}$ of the arms to exist such that $p_{ij} > \frac{1}{2}$ for all $i < j$. Here we provide evidence that this assumption is often violated in practice. By contrast, the algorithm we propose in Section 4 makes only the Condorcet assumption, which is implied by the total ordering assumption of IF and BTM.

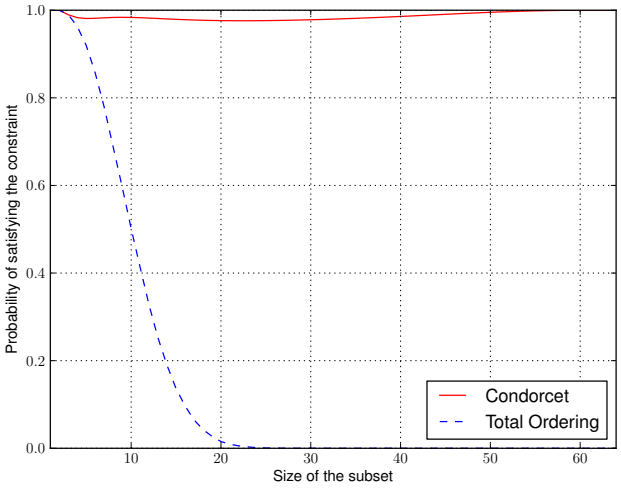

Figure 3. The probability that the Condorcet and the total ordering assumptions hold for subsets of the feature rankers. The probability is shown as a function of the size of the subset.

In order to test how stringent an assumption the existence of a Condorcet winner is compared to the total ordering assumption, we estimated the probability of each assumption holding in our ranker evaluation application. Using the same preference matrix as in our experiments in Section 6, we computed for each $K = 1, \ldots, 64$ the probability P_K that a given K-armed dueling bandit problem obtained from considering K of our 64 feature rankers would have a Condorcet winner as follows: first, we calculated the number of K-armed dueling bandit problems that have a Condorcet winner by calculating for each feature ranker r how many K-armed dueling bandit problems it can be the Condorcet winner of: for each r, this is equal to $\binom{N_r}{K}$, where N_r is the number rankers that r beats; next, we divided this total number of K-armed dueling bandit problems with a Condorcet winner by $\binom{64}{K}$, which is the number of all K-armed dueling bandit problems that one could construct from these 64 rankers.

The probabilities P_K, plotted as a function of K in Figure 3 (the red curve), were all larger than 0.97. The same plot also shows an estimate of the probability that the total ordering assumption holds for a given K (the blue curve), which was obtained by randomly selecting $100,000$ K-armed dueling bandit problems and searching for ones that satisfy the total ordering assumption. As can be seen from Figure 3, as K grows the probability that the total ordering assumption holds decreases rapidly. This is because there exist cyclical relationships between these feature rankers and as soon as the chosen subset of feature rankers contains one of these cycles, it fails to satisfy the total ordering condition. By contrast, the Condorcet assumption will still be satisfied as long as the cycle does not include the Condorcet winner. Moreover, because of the presence of these cycles, the probability that the Condorcet assumption holds decreases initially as K increases, but then increases again because the number of all possible K-armed dueling bandit decreases as K approaches 64.

Furthermore, in addition to the total ordering assumption, IF and BTM each require a form of *stochastic transitivity*. In particular, IF requires *strong stochastic transitivity*; for any triple (i, j, k), with $i < j < k$, the following condition needs to be satisfied:

$$p_{ik} \geq \max\{p_{ij}, p_{jk}\}.$$

BTM requires the less restrictive *relaxed stochastic transi-*

tivity, i.e., that there exists a number $\gamma \geq 1$ such that for all pairs (j, k) with $1 < j < k$, we have

$$\gamma p_{1k} \geq \max\{p_{1j}, p_{jk}\}.$$

As pointed out in (Yue & Joachims, 2011), strong stochastic transitivity is often violated in practice, a phenomenon also observed in our experiments: for instance, all of the K-armed dueling bandit problems on which we experimented require $\gamma > 1$.

Even though BTM permits a broader class of K-armed dueling bandit problems, it requires γ to be explicitly passed to it as a parameter, which poses substantial difficulties in practice. If γ is underestimated, the algorithm can in certain circumstances be misled with high probability into choosing the Borda winner instead of the Condorcet winner. On the other hand, though overestimating γ does not cause the algorithm to choose the wrong arm, it nonetheless results in a severe penalty, since it makes the algorithm much more exploratory, yielding the γ^7 term in the upper bound on the cumulative regret, as discussed in Section 3. For instance, in the three-ranker evaluation experiments discussed in Section 6, the values for γ are 4.85, 11.6 and 47.3 for the 16-, 32- and 64-armed examples: even the smallest of these numbers raised to the power of 7 is on the order of tens of thousands, making this upper bound very large.

8.2. Proof of Lemma 1

In this section, we prove Lemma 1, whose statement is repeated here for convenience. Recall from Section 5 that we assume without loss of generality that a_1 is the optimal arm. Moreover, given any K-armed dueling bandit algorithm, we define $w_{ij}(t)$ to be the number of times arm a_i has beaten a_j in the first t iterations of the algorithm. We also define $u_{ij}(t) := \frac{w_{ij}(t)}{w_{ij}(t)+w_{ji}(t)} + \sqrt{\frac{\alpha \ln t}{w_{ij}(t)+w_{ji}(t)}}$, where α is any positive contant, and $l_{ij}(t) := 1 - u_{ji}(t)$. Moreover, for any $\delta > 0$, define $C(\delta) := \left(\frac{(4\alpha-1)K^2}{(2\alpha-1)\delta}\right)^{\frac{1}{2\alpha-1}}$.

Lemma 1. *Let $\mathbf{P} := [p_{ij}]$ be the preference matrix of a K-armed dueling bandit problem with arms $\{a_1, \ldots, a_K\}$. Then, for any dueling bandit algorithm and any $\alpha > \frac{1}{2}$ and $\delta > 0$, we have*

$$P\Big(\forall\, t > C(\delta), i, j,\ p_{ij} \in [l_{ij}(t), u_{ij}(t)]\Big) > 1 - \delta. \quad (9)$$

Proof. To decompose the lefthand side of (9), we introduce the notation $\mathcal{G}_{ij}(t)$ for the "good" event that at time t we have $p_{ij} \in [l_{ij}(t), u_{ij}(t)]$, which satisfies the following:

(i) $\mathcal{G}_{ij}(t) = \mathcal{G}_{ji}(t)$ because of the triple of equalities $\Big(p_{ji}, l_{ji}(t), u_{ji}(t)\Big) = \Big(1 - p_{ij}, 1 - u_{ij}(t), 1 - l_{ij}(t)\Big)$.

(ii) $\mathcal{G}_{ii}(t)$ always holds, since $(p_{ii}, l_{ii}(t), u_{ii}(t)) = \left(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}\right)$. Together with (i), this means that we only need to consider $\mathcal{G}_{ij}(t)$ for $i < j$.

(iii) Define τ_n^{ij} to be the iteration at which arms i and j were compared against each other for the n^{th} time. If $G_{ij}\left(\tau_n^{ij} + 1\right)$ holds, then the events $\mathcal{G}_{ij}(t)$ hold for all $t \in \left(\tau_n^{ij}, \tau_{n+1}^{ij}\right]$ because when $t \in \left(\tau_n^{ij}, \tau_{n+1}^{ij}\right]$, w_{ij} and w_{ji} remain constant and so in the expressions for $u_{ij}(t)$ and $u_{ji}(t)$ only the $\ln t$ changes, which is a monotonically increasing function of t. So, we have

$$l_{ij}(t) \leq l_{ij}(\tau_n^{ij} + 1) \leq p_{ij} \leq u_{ij}(\tau_n^{ij} + 1) \leq u_{ij}(t).$$

Moreover, the same statement holds with τ_n^{ij} replaced by any $T \in \left(\tau_n^{ij}, \tau_{n+1}^{ij}\right]$, i.e., if we know that $\mathcal{G}_{ij}(T)$ holds, then $\mathcal{G}_{ij}(t)$ also holds for all $t \in \left(T, \tau_{n+1}^{ij}\right]$. This is illustrated in Figure 4.

Now, given the above three facts, we have for any T

$$\begin{aligned} &P\Big(\forall\, t \geq T, i, j,\ \mathcal{G}_{ij}(t)\Big) \quad (10)\\ &= P\Big(\forall\, i > j,\ \mathcal{G}_{ij}(T) \text{ and } \forall\, n\ s.t.\ \tau_n^{ij} > T,\ \mathcal{G}_{ij}(\tau_n^{ij})\Big). \end{aligned}$$

Let us now flip things around and look at the complement of these events, i.e. the "bad" event $\mathcal{B}_{ij}(t)$ that

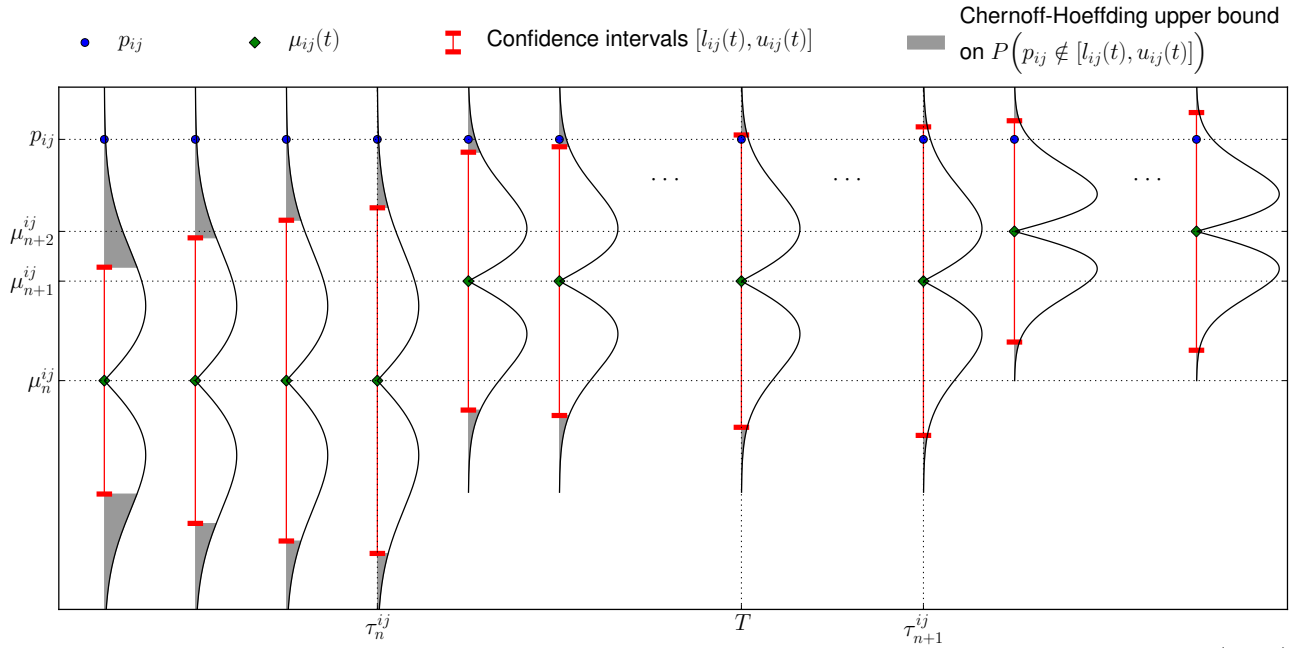

Figure 4. An illustrations of the idea behind Lemma 1 using an example of how the confidence intervals of a single pair of arms (a_i, a_j), and their relation to the comparison probability p_{ij}, might evolve over time. The time-step τ^{ij}_m denotes the m^{th} time when the arms a_i and a_j were chosen by RUCB to be compared against each other. We also define $\mu^{ij}_m := \mu_{ij}(\tau^{ij}_m)$. The time T is when the confidence intervals $[l_{ij}(t), u_{ij}(t)]$ begin to include p_{ij}. The lemma then states that with probability $1-\delta$, we have $T \leq C(\delta)$.

Moreover, for each time-step, the area of the shaded region under the vertical graphs is the bound given by the Chernoff-Hoeffding (CH) bound on the probability that the confidence interval will not contain p_{ij}. Note that the CH bound has the form $e^{-(x-\mu^{ij}_n)^2}$ and so in order for this number to be the area under a graph (hence making it easier to illustrate in a figure), we have drawn the derivative of this function, $f^{ij}_n(x) := |x-\mu^{ij}_n| e^{-(x-\mu^{ij}_n)^2}$, which is why the graphs are equal to 0 in the middle. Note that this does not mean that μ^{ij}_n has very low probability of being close to p_{ij}: the graphs drawn here are not the PDFs of the posteriors, but simply a manifestation of the bound given by the Chernoff-Hoeffding bound. More specifically, the property that they satisfy is that $P\Big(p_{ij} \notin [l_{ij}(t), u_{ij}(t)]\Big) \leq \int_{-\infty}^{l_{ij}(t)} f^{ij}_{N_{ij}(t)}(x)dx + \int_{u_{ij}(t)}^{\infty} f^{ij}_{N_{ij}(t)}(x)dx$.

$p_{ij} \notin [l_{ij}(t), u_{ij}(t)]$ occurs. Subtracting both sides of Equation (10) from 1 and using the union bound gives

$$P\Big(\exists t > T, i, j \; s.t. \; \mathcal{B}_{ij}(t)\Big)$$
$$\leq \sum_{i<j} \left[P\Big(\mathcal{B}_{ij}(T)\Big) + P\Big(\exists n : \tau^{ij}_n > T \text{ and } \mathcal{B}_{ij}(\tau^{ij}_n)\Big) \right].$$

Further decomposing the righthand side using union bounds and making the condition explicit, we get

$$P\Big(\exists t > T, i, j \; s.t. \; \mathcal{B}_{ij}(t)\Big)$$
$$\leq \sum_{i>j} \left[P\left(\left| p_{ij} - \mu^{ij}_{N_{ij}(T)} \right| > \sqrt{\frac{\alpha \ln T}{N_{ij}(T)}} \right) + \right.$$
$$P\left(\exists n \leq T \; s.t. \; \tau^{ij}_n > T \text{ and } \left| p_{ij} - \mu^{ij}_n \right| > \sqrt{\frac{\alpha \ln \tau^{ij}_n}{n}} \right)$$
$$\left. + P\left(\exists n > T \; s.t. \; \left| p_{ij} - \mu^{ij}_n \right| > \sqrt{\frac{\alpha \ln \tau^{ij}_n}{n}} \right) \right],$$

since $T < n < \tau^{ij}_n$. Here, $\mu^{ij}_n := \frac{w_{ij}(\tau^{ij}_n)}{w_{ij}(\tau^{ij}_n) + w_{ji}(\tau^{ij}_n)}$ is the frequentist estimate of p_{ij} after n comparisons between arms a_i and a_j.

Now, in the above sum, we can upper-bound the first term by looking at the higher probability event that $\mathcal{B}_{ij}(T)$ happens for any possible number of comparisons between a_i and a_j, and since we know that $N_{ij}(T) \leq T$, we can replace $N_{ij}(T)$ with a variable n that can take values between 0 and T. For the second term, we know that $\tau^{ij}_n > T$, so we can replace τ^{ij}_n with T and remove the condition $\tau^{ij}_n > T$ and look at all $n \leq T$. For the third term, since we always have that $n < \tau^{ij}_n$, we can replace τ^{ij}_n with n and get a higher probability event. Putting all of this together, we get the following looser bound:

$$
\begin{aligned}
&P\Big(\exists t > T, i, j \ s.t. \ \mathcal{B}_{ij}(t)\Big) \\
&\leq \sum_{i<j} \Bigg[P\left(\exists n \in \{0, \dots, T\} : \left|p_{ij} - \mu_n^{ij}\right| > \sqrt{\frac{\alpha \ln T}{n}}\right) \\
&\qquad + P\left(\exists n \in \{0, \dots, T\} : \left|p_{ij} - \mu_n^{ij}\right| > \sqrt{\frac{\alpha \ln T}{n}}\right) \\
&\qquad + P\left(\exists n > T \ s.t. \ \left|p_{ij} - \mu_n^{ij}\right| > \sqrt{\frac{\alpha \ln n}{n}}\right) \Bigg] \\
&\leq \sum_{i<j} \Bigg[2 \sum_{n=0}^{T} P\left(\left|p_{ij} - \mu_n^{ij}\right| > \sqrt{\frac{\alpha \ln T}{n}}\right) \\
&\qquad + \sum_{n=T+1}^{\infty} P\left(\left|p_{ij} - \mu_n^{ij}\right| > \sqrt{\frac{\alpha \ln n}{n}}\right) \Bigg]. \qquad (11)
\end{aligned}
$$

To bound the expression on line (11), we apply the Chernoff-Hoeffding bound, which in its simplest form states that given i.i.d. random variables $X_1, \dots, X_n$, whose support is contained in $[0, 1]$ and whose expectation satisfies $\mathbb{E}[X_k] = p$, and defining $\mu_n := \frac{X_1 + \cdots + X_n}{n}$, we have $P(|\mu_n - p| > a) \leq 2e^{-2na^2}$. This gives us

$$
\begin{aligned}
&P\Big(\exists t > T, i, j \ s.t. \ \mathcal{B}_{ij}(t)\Big) \\
&\leq \sum_{i<j} \left[2 \sum_{n=1}^{T} 2e^{-2\cancel{n}\frac{\alpha \ln T}{\cancel{n}}} + \sum_{n=T+1}^{\infty} 2e^{-2\cancel{n}\frac{\alpha \ln n}{\cancel{n}}} \right] \\
&= \frac{K(K-1)}{2} \left[\sum_{n=1}^{T} \frac{4}{T^{2\alpha}} + \sum_{n=T+1}^{\infty} \frac{2}{n^{2\alpha}} \right] \\
&\leq \frac{2K^2}{T^{2\alpha-1}} + K^2 \int_T^{\infty} \frac{dx}{x^{2\alpha}}, \text{ since } \frac{1}{x^{2\alpha}} \text{ is decreasing.} \\
&\leq \frac{2K^2}{T^{2\alpha-1}} + K^2 \int_T^{\infty} \frac{dx}{x^{2\alpha}} \\
&= \frac{2K^2}{T^{2\alpha-1}} + \frac{K^2}{(1-2\alpha)x^{2\alpha-1}} \Bigg|_T^{\infty} \\
&= \frac{(4\alpha-1)K^2}{(2\alpha-1)T^{2\alpha-1}}. \qquad (12)
\end{aligned}
$$

Now, since $C(\delta) = \left(\frac{(4\alpha-1)K^2}{(2\alpha-1)\delta}\right)^{\frac{1}{2\alpha-1}}$ for each $\delta > 0$, the bound in (12) gives us (9).

□

8.3. Proof of Theorem 5

Here, we provide the proof of the expected regret bound claimed in Theorem 5, starting by repeating the statement of the theorem:

Theorem 5. *Given the setup of Proposition 2 together with the notation of Theorem 4, we have the following expected regret bound for RUCB, where the expectations are taken across different runs of the algorithm: if we have $\alpha > 1$, the expected regret accumulated by RUCB after T iterations is bounded by*

$$
\begin{aligned}
\mathbb{E}[R_T] \ \leq \ &\left[8 + \left(\frac{2(4\alpha-1)K^2}{2\alpha-1}\right)^{\frac{1}{2\alpha-1}} \frac{2\alpha-1}{\alpha-1} \right] \Delta_{\max} \\
&+ 2D\Delta_{\max} \ln 2D + \sum_{j=2}^{K} \frac{2\alpha\left(\Delta_j + 4\Delta_{\max}\right)}{\Delta_j^2} \ln T. \qquad (13)
\end{aligned}
$$

Proof. We can obtain the bound in (13) from (6) by integrating with respect to δ from 0 to 1. This is because given any one-dimensional random variable X with CDF F_X, we can use the identity $\mathbb{E}[X] = \int_0^1 F_X^{-1}(q)dq$. In our case, $X = R_T$ for a fixed time t and, as illustrated in Figure 5, we can deduce from (6) that $F_{R_T}(r) > H_T^{-1}(r)$, which gives the bound

$$
F_{R_T}^{-1}(q) < H_T(q) = \widehat{C}(1-q) + \sum_{j=2}^{K} \widehat{D}_j \ln T.
$$

Now, assume that $\alpha > 1$. To derive (13) from the above inequality, we need to integrate the righthand side, and since it is only the first two terms in the definition of $\widehat{C}$ that depends on δ, that is all we need to integrate. Let us deal

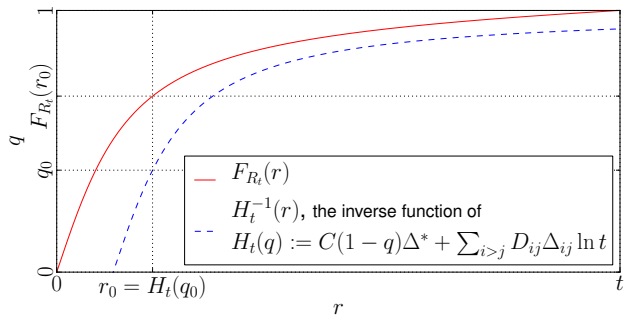

Figure 5. A schematic graph illustrating the proof of Theorem 5. Note that the expression for $H_T(q)$ is extracted from (6), which also implies that H_T^{-1} is necessarily below F_{R_T}: formulated in terms of CDFs, (6) states that $F_{R_T}\left(H_T(q_0)\right) > q_0 = H_T^{-1}\left(H_T(q_0)\right)$, where $q_0 = 1 - \delta_0$ is a quantile. From this, we can conclude that $F_{R_T}(r) > H_T^{-1}(r)$ for all r.

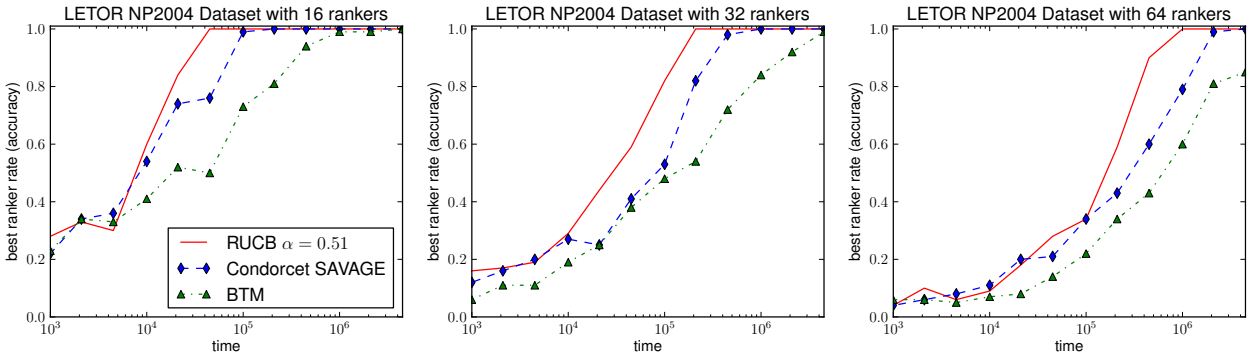

Figure 6. Average accuracy for 100 runs of BTM, Condorcet SAVAGE and RUCB with $\alpha = 0.51$ applied to three K-armed dueling bandit problems with $K = 16, 32, 64$. Note that the x-axes in these plots use a log scale.

with the first term first, using the substitution $1 - q = \delta$, $dq = -d\delta$:

$$\begin{aligned}\int_{q=0}^{1} 4\Delta_{\max} \ln \frac{2}{1-q} dq &= 4\Delta_{\max}\left[\ln 2 - \int_{\delta=1}^{0} -\ln \delta \, d\delta\right] \\ &= 4\Delta_{\max}\left[\ln 2 - \int_{\delta=0}^{1} \ln \delta \, d\delta\right] \\ &= 4\Delta_{\max}(\ln 2 + 1) < 8\Delta_{\max}\end{aligned}$$

To deal with the second term in $\widehat{C}$, recall that it is equal to $2\Delta_{\max} C\left(\frac{\delta}{2}\right) := 2\Delta_{\max}\left(\frac{2(4\alpha-1)K^2}{(2\alpha-1)\delta}\right)^{\frac{1}{2\alpha-1}}$, so to simplify notation, we define

$$L := 2\Delta_{\max}\left(\frac{2(4\alpha-1)K^2}{2\alpha-1}\right)^{\frac{1}{2\alpha-1}}.$$

Now, we can carry out the integration as follows, again using the substitution $1 - q = \delta$, $dq = -d\delta$:

$$\begin{aligned}\int_{q=0}^{1} C(1-q)dq &= \int_{\delta=1}^{0} -C(\delta)d\delta \\ &= \int_{0}^{1} 2\left(\frac{2(4\alpha-1)K^2}{(2\alpha-1)\delta}\right)^{\frac{1}{2\alpha-1}} d\delta \\ &= L\int_{0}^{1} \delta^{-\frac{1}{2\alpha-1}} d\delta \\ &= L\left[\frac{\delta^{1-\frac{1}{2\alpha-1}}}{1-\frac{1}{2\alpha-1}}\right]_0^1 \\ &= \left(\frac{2(4\alpha-1)K^2}{2\alpha-1}\right)^{\frac{1}{2\alpha-1}} \frac{2\alpha-1}{\alpha-1}.\end{aligned}$$

□

8.4. Experimental Details

Our experimental setup is built on real IR data, namely the LETOR NP2004 dataset (Liu et al., 2007). This dataset is based on the TREC Web track named-page finding task, where a query is what the user believes to be a reasonable estimate of the name of the webpage she is seeking. Using this data set, we create a set of 64 rankers, each corresponding to a ranking feature provided in the data set, e.g., PageRank. The ranker evaluation task in this context corresponds to determining which single feature constitutes the best ranker (Hofmann et al., 2013a).

To compare a pair of rankers, we use *probabilistic interleave* (PI) (Hofmann et al., 2011), a recently developed method for interleaved comparisons. To model the user's click behavior on the resulting interleaved lists, we employ a probabilistic user model (Hofmann et al., 2011; Craswell et al., 2008) that uses as input the manual labels (classifying documents as relevant or not for given queries) provided with the LETOR NP2004 dataset. Queries are sampled randomly and clicks are generated probabilistically by conditioning on these assessments in a way that resembles the behavior of an actual user (Guo et al., 2009).

Following (Yue & Joachims, 2011), we first used the above approach to estimate the comparison probabilities p_{ij} for each pair of rankers and then used these probabilities to simulate comparisons between rankers. More specifically, we estimated the full preference matrix by performing 4000 interleaved comparisons on each pair of the 64 feature rankers.

Finally, the plots in Figure 6 show the accuracy of all three algorithms across 100 runs, computed at the same times as the exploration horizons used for BTM and SAVAGE in Figure 2. Note that RUCB reaches the 80% mark almost twice as fast as Condorcet SAVAGE, all without knowing the horizon T. The contrast is even more stark when comparing to BTM.