Dueling Bandits: Beyond Condorcet Winners to General Tournament Solutions

Siddartha Ramamohan
Indian Institute of Science
Bangalore 560012, India
siddartha.yr@csa.iisc.ernet.in

Arun Rajkumar
Xerox Research
Bangalore 560103, India
arun_r@csa.iisc.ernet.in

Shivani Agarwal
University of Pennsylvania
Philadelphia, PA 19104, USA
ashivani@seas.upenn.edu

Abstract

Recent work on deriving $O(\log T)$ anytime regret bounds for stochastic dueling bandit problems has considered mostly Condorcet winners, which do not always exist, and more recently, winners defined by the Copeland set, which do always exist. In this work, we consider a broad notion of winners defined by tournament solutions in social choice theory, which include the Copeland set as a special case but also include several other notions of winners such as the top cycle, uncovered set, and Banks set, and which, like the Copeland set, always exist. We develop a family of UCB-style dueling bandit algorithms for such general tournament solutions, and show $O(\log T)$ anytime regret bounds for them. Experiments confirm the ability of our algorithms to achieve low regret relative to the target winning set of interest.

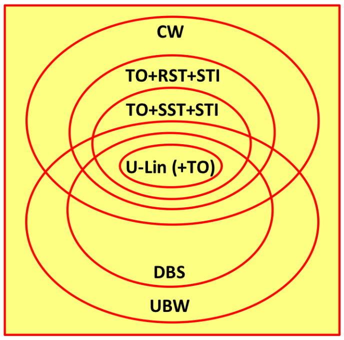
1 Introduction

There has been significant interest and progress in recent years in developing algorithms for dueling bandit problems [1–11]. Here there are K arms; on each trial t, one selects a pair of arms (i_t, j_t) for comparison, and receives a binary feedback signal $y_t \in \{0, 1\}$ indicating which arm was preferred. Most work on dueling bandits is in the stochastic setting and assumes a stochastic model – a preference matrix $\mathbf{P}$ of pairwise comparison probabilities P_{ij} – from which the feedback signals y_t are drawn; as with standard stochastic multi-armed bandits, the target here is usually to design algorithms with $O(\ln T)$ regret bounds, and where possible, $O(\ln T)$ *anytime* (or 'horizon-free') regret bounds, for which the algorithm does not need to know the horizon or number of trials T in advance.

Early work on dueling bandits often assumed strong conditions on the preference matrix $\mathbf{P}$, such as existence of a total order, under which there is a natural notion of a 'maximal' element with respect to which regret is measured. Recent work has sought to design algorithms under weaker conditions on $\mathbf{P}$; most work, however, has assumed the existence of a Condorcet winner, which is an arm i that beats every other arm j ($P_{ij} > \frac{1}{2}\ \forall j \neq i$), and which reduces to the maximal element when a total order exists. Unfortunately, the Condorcet winner does not always exist, and this has motivated a search for other natural notions of winners, such as Borda winners and the Copeland set (see Figure 1).[1] Among these, the only work that offers anytime $O(\ln T)$ regret bounds is the recent work of Zoghi et al. [11] on Copeland sets. In this work, we consider defining winners in dueling bandits via the natural notion of *tournament solutions* used in social choice theory, of which the Copeland set is a special case. We develop general upper confidence bound (UCB) style dueling bandit algorithms for a number of tournament solutions including the top cycle, uncovered set, and Banks set, and prove $O(\ln T)$ anytime regret bounds for them, where the regret is measured relative to the tournament solution of interest. Our proof technique is modular and can be used to develop algorithms with similar bounds for any tournament solution for which a 'selection procedure' satisfying certain 'safety conditions' can be designed. Experiments confirm the ability of our algorithms to achieve low regret relative to the target winning set of interest.

[1] Recently, Dudik et al. [10] also studied von Neumann winners, although they did so in a different (contextual) setting, leading to $O(T^{1/2})$ and $O(T^{2/3})$ regret bounds.

29th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.



Algorithm	Condition on P	Target Winner	Anytime?
MultiSBM [5]	U-Lin	Condorcet winner	✓
IF [1]	TO+SST+STI	Condorcet winner	×
BTMB [2]	TO+RST+STI	Condorcet winner	×
RUCB [6]	CW	Condorcet winner	✓
MergeRUCB [7]	CW	Condorcet winner	✓
RMED [9]	CW	Condorcet winner	✓
SECS [8]	UBW	Borda winner	×
PBR-SE [4]	DBS	Borda winner	×
PBR-CO [4]	Any $\mathbf{P}$ without ties	Copeland set	×
SAVAGE-BO [3]	Any $\mathbf{P}$ without ties	Borda winner	×
SAVAGE-CO [3]	Any $\mathbf{P}$ without ties	Copeland set	×
CCB, SCB [11]	Any $\mathbf{P}$ without ties	Copeland set	✓
UCB-TC	Any $\mathbf{P}$ without ties	Top cycle	✓
UCB-UC	Any $\mathbf{P}$ without ties	Uncovered set	✓
UCB-BA	Any $\mathbf{P}$ without ties	Banks set	✓

Figure 1: Summary of algorithms for stochastic dueling bandit problems that have $O(\ln T)$ regret bounds, together with corresponding conditions on the underlying preference matrix $\mathbf{P}$, target winners used in defining regret, and whether the regret bounds are "anytime". The figure on the left shows relations between some of the commonly studied conditions on $\mathbf{P}$ (see Table 1 for definitions). The algorithms in the lower part of the table (shown in red) are proposed in this paper.

2 Dueling Bandits, Tournament Solutions, and Regret Measures

Dueling Bandits. We denote by $[K] = \{1, \ldots, K\}$ the set of K arms. On each trial t, the learner selects a pair of arms $(i_t, j_t) \in [K] \times [K]$ (with i_t possibly equal to j_t), and receives feedback in the form of a comparison outcome $y_t \in \{0, 1\}$, with $y_t = 1$ indicating i_t was preferred over j_t and $y_t = 0$ indicating the reverse. The goal of the learner is to select as often as possible from a set of 'good' or 'winning' arms, which we formalize below as a tournament solution.

The pairwise feedback on each trial is assumed to be generated stochastically according to a fixed but unknown pairwise preference model represented by a *preference matrix* $\mathbf{P} \in [0, 1]^{K \times K}$ with $P_{ij} + P_{ji} = 1 \; \forall i, j$: whenever arms i and j are compared, i is preferred to j with probability P_{ij}, and j to i with probability $P_{ji} = 1 - P_{ij}$. Thus for each trial t, we have $y_t \sim \text{Bernoulli}(P_{i_t j_t})$. We assume throughout this paper that there are are no "ties" between distinct arms, i.e. that $P_{ij} \neq \frac{1}{2} \; \forall i \neq j$.[2] We denote by $\mathcal{P}_K$ the set of all such preference matrices over K arms:

$$\mathcal{P}_K \;\; = \;\; \left\{ \mathbf{P} \in [0, 1]^{K \times K} : P_{ij} + P_{ji} = 1 \; \forall i, j \, ; \; P_{ij} \neq \tfrac{1}{2} \; \forall i \neq j \right\}.$$

For any pair of arms (i, j), we will define the margin of (i, j) w.r.t. $\mathbf{P}$ as

$$\Delta^{\mathbf{P}}_{ij} = |P_{ij} - \tfrac{1}{2}|\,.$$

Previous work on dueling bandits has considered a variety of conditions on $\mathbf{P}$; see Table 1 and Figure 1. Our interest here is in designing algorithms that have regret guarantees under minimal restrictions on $\mathbf{P}$. To this end, we will consider general notions of winners that are derived from a natural tournament associated with $\mathbf{P}$, and that are always guaranteed to exist. We will say an arm i *beats* an arm j w.r.t. $\mathbf{P}$ if $P_{ij} > \frac{1}{2}$; we will express this as a binary relation $\succ_{\mathbf{P}}$ on $[K]$:

$$i \succ_{\mathbf{P}} j \iff P_{ij} > \tfrac{1}{2}\,.$$

The *tournament associated with* $\mathbf{P}$ is then simply $\mathcal{T}_{\mathbf{P}} = ([K], E_{\mathbf{P}})$, where $E_{\mathbf{P}} = \{(i, j) : i \succ_{\mathbf{P}} j\}$. Two frequently studied notions of winners in previous work on dueling bandits, both of which are derived from the tournament $\mathcal{T}_{\mathbf{P}}$ (and which are the targets of previous anytime regret bounds), are the Condorcet winner when it exists, and the Copeland set in general:

Definition 1 (Condorcet winner). *Let* $\mathbf{P} \in \mathcal{P}_K$. *If there exists an arm* $i^* \in [K]$ *such that* $i^* \succ_{\mathbf{P}} j \; \forall j \neq i^*$, *then* i^* *is said to be a* Condorcet winner *w.r.t.* $\mathbf{P}$.

Definition 2 (Copeland set). *Let* $\mathbf{P} \in \mathcal{P}_K$. *The* Copeland set *w.r.t.* $\mathbf{P}$, *denoted* $\text{CO}(\mathbf{P})$, *is defined as the set of all arms in* $[K]$ *that beat the maximal number of arms w.r.t.* $\mathbf{P}$:

$$\text{CO}(\mathbf{P}) = \arg\max_{i \in [K]} \textstyle\sum_{j \neq i} \mathbf{1}\big(i \succ_{\mathbf{P}} j\big)\,.$$

Here we are interested in more general notions of winning sets derived from the tournament $\mathcal{T}_{\mathbf{P}}$.

[2] The assumption of no ties was also made in deriving regret bounds w.r.t. to the Copeland set in [3, 4, 11], and exists implicitly in [1, 2] as well.

Table 1: Commonly studied conditions on the preference matrix $\mathbf{P}$.

Condition on $\mathbf{P}$	**Property satisfied by $\mathbf{P}$**
Utility-based with linear link (U-Lin)	$\exists \mathbf{u} \in [0,1]^K : P_{ij} = \frac{1-(u_i-u_j)}{2} \; \forall i,j$
Total order (TO)	$\exists \sigma \in \mathcal{S}_n : P_{ij} > \frac{1}{2} \iff \sigma(i) < \sigma(j)$
Strong stochastic transitivity (SST)	$P_{ij} > \frac{1}{2}, P_{jk} > \frac{1}{2} \implies P_{ik} \geq \max(P_{ij}, P_{jk})$
Relaxed stochastic transitivity (RST)	$\exists \gamma \geq 1 : P_{ij} > \frac{1}{2}, P_{jk} > \frac{1}{2} \implies P_{ik} - \frac{1}{2} \geq \frac{1}{\gamma} \max(P_{ij} - \frac{1}{2}, P_{jk} - \frac{1}{2})$
Stochastic triangle inequality (STI)	$P_{ij} > \frac{1}{2}, P_{jk} > \frac{1}{2} \implies P_{ik} \leq P_{ij} + P_{jk} - \frac{1}{2}$
Condorcet winner (CW)	$\exists i : P_{ij} > \frac{1}{2} \; \forall j \neq i$
Unique Borda winner (UBW)	$\exists i : \sum_{k \neq i} P_{ik} > \sum_{k \neq j} P_{jk} \; \forall j \neq i$
Distinct Borda scores (DBS)	$\sum_{k \neq i} P_{ik} \neq \sum_{k \neq j} P_{jk} \; \forall i \neq j$

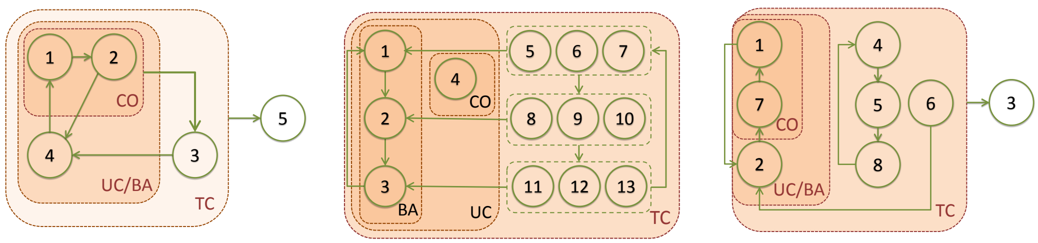


Figure 2: Examples of various tournaments together with their corresponding tournament solutions. Edges that are not explicitly shown are directed from left to right; edges that are incident on subsets of nodes (rounded rectangles) apply to all nodes within. **Left:** A tournament on 5 nodes with gradually discriminating tournament solutions. **Middle:** The Hudry tournament on 13 nodes with disjoint Copeland and Banks sets. **Right:** A tournament on 8 nodes based on ATP tennis match records.

Tournament Solutions. Tournament solutions have long been used in social choice and voting theory to define winners in general tournaments when no Condorcet winner exists [12, 13]. Specifically, a tournament solution is any mapping that maps each tournament on K nodes to a subset of 'winning' nodes in $[K]$; for our purposes, we will define a tournament solution to be any mapping $\mathsf{S} : \mathcal{P}_K \to 2^{[K]}$ that maps each preference matrix $\mathbf{P}$ (via the induced tournament $\mathcal{T}_{\mathbf{P}}$) to a subset of winning arms $\mathsf{S}(\mathbf{P}) \subseteq [K]$.[3] The Copeland set is one such tournament solution. We consider three additional tournament solutions in this paper: the top cycle, the uncovered set, and the Banks set, all of which offer other natural generalizations of the Condorcet winner. These tournament solutions are motivated by different considerations (ranging from dominance to covering to decomposition into acyclic subtournaments) and have graded discriminative power, and can therefore be used to match the needs of different applications; see [12] for a comprehensive survey.

Definition 3 (Top cycle). *Let $\mathbf{P} \in \mathcal{P}_K$. The* top cycle *w.r.t. $\mathbf{P}$, denoted* $\mathrm{TC}(\mathbf{P})$*, is defined as the smallest set $W \subseteq [K]$ for which $i \succ_{\mathbf{P}} j \; \forall i \in W, j \notin W$.*

Definition 4 (Uncovered set). *Let $\mathbf{P} \in \mathcal{P}_K$. An arm i is said to* cover *an arm j w.r.t. $\mathbf{P}$ if $i \succ_{\mathbf{P}} j$ and $\forall k : j \succ_{\mathbf{P}} k \implies i \succ_{\mathbf{P}} k$. The* uncovered set *w.r.t. $\mathbf{P}$, denoted* $\mathrm{UC}(\mathbf{P})$*, is defined as the set of all arms that are not covered by any other arm w.r.t. $\mathbf{P}$:*

$$\mathrm{UC}(\mathbf{P}) = \left\{ i \in [K] : \nexists j \in [K] \text{ s.t. } j \text{ covers } i \text{ w.r.t. } \mathbf{P} \right\}.$$

Definition 5 (Banks set). *Let $\mathbf{P} \in \mathcal{P}_K$. A subtournament $\mathcal{T} = (V, E)$ of $\mathcal{T}_{\mathbf{P}}$, where $V \subseteq [K]$ and $E = E_{\mathbf{P}}|_{V \times V}$, is said to be* maximal acyclic *if (i) $\mathcal{T}$ is acyclic, and (ii) no other subtournament containing $\mathcal{T}$ is acyclic. Denote by* $\mathrm{MAST}(\mathbf{P})$ *the set of all maximal acyclic subtournaments of $\mathcal{T}_{\mathbf{P}}$, and for each $\mathcal{T} \in \mathrm{MAST}(\mathbf{P})$, denote by $m^*(\mathcal{T})$ the maximal element of $\mathcal{T}$. Then the* Banks set *w.r.t. $\mathbf{P}$, denoted* $\mathrm{BA}(\mathbf{P})$*, is defined as the set of maximal elements of all maximal acyclic subtournaments of $\mathcal{T}_{\mathbf{P}}$:*

$$\mathrm{BA}(\mathbf{P}) = \left\{ m^*(\mathcal{T}) : \mathcal{T} \in \mathrm{MAST}(\mathbf{P}) \right\}.$$

It is known that $\mathrm{BA}(\mathbf{P}) \subseteq \mathrm{UC}(\mathbf{P}) \subseteq \mathrm{TC}(\mathbf{P})$ and $\mathrm{CO}(\mathbf{P}) \subseteq \mathrm{UC}(\mathbf{P}) \subseteq \mathrm{TC}(\mathbf{P})$. In general, $\mathrm{BA}(\mathbf{P})$ and $\mathrm{CO}(\mathbf{P})$ may intersect, although they can also be disjoint. When $\mathbf{P}$ contains a Condorcet winner i^*, all four tournament solutions reduce to just the singleton set $\{i^*\}$. See Figure 2 for examples.

[3] Strictly speaking, the mapping S must be invariant under permutations of the node labels.

Regret Measures. When $\mathbf{P}$ admits a Condorcet winner i^*, the individual regret of an arm i is usually defined as $r^{\text{CW}}_{\mathbf{P}}(i) = \Delta^{\mathbf{P}}_{i^*,i}$, and the cumulative regret over T trials of an algorithm $\mathcal{A}$ that selects arms (i_t, j_t) on trial t is then generally defined as $\mathcal{R}^{\text{CW}}_T(\mathcal{A}) = \sum_{t=1}^{T} r^{\text{CW}}_{\mathbf{P}}(i_t, j_t)$, where the pairwise regret $r^{\text{CW}}_{\mathbf{P}}(i,j)$ is either the average regret $\frac{1}{2}\big(r^{\text{CW}}_{\mathbf{P}}(i) + r^{\text{CW}}_{\mathbf{P}}(j)\big)$, the strong regret $\max\big(r^{\text{CW}}_{\mathbf{P}}(i), r^{\text{CW}}_{\mathbf{P}}(j)\big)$, or the weak regret $\min\big(r^{\text{CW}}_{\mathbf{P}}(i), r^{\text{CW}}_{\mathbf{P}}(j)\big)$ [1, 2, 6, 7, 9].[4] When the target winner is a tournament solution S, we can similarly define a suitable notion of individual regret of an arm i w.r.t. S, and then use this to define pairwise regrets as above.

In particular, for the three tournament solutions discussed above, we will define the following natural notions of individual regret:

$$r^{\text{TC}}_{\mathbf{P}}(i) = \begin{cases} \max\limits_{i^* \in \text{TC}(\mathbf{P})} \Delta^{\mathbf{P}}_{i^*,i} & \text{if } i \notin \text{TC}(\mathbf{P}) \\ 0 & \text{if } i \in \text{TC}(\mathbf{P}) \end{cases}; \qquad r^{\text{UC}}_{\mathbf{P}}(i) = \begin{cases} \max\limits_{i^* \in \text{UC}(\mathbf{P}): i^* \text{ covers } i} \Delta^{\mathbf{P}}_{i^*,i} & \text{if } i \notin \text{UC}(\mathbf{P}) \\ 0 & \text{if } i \in \text{UC}(\mathbf{P}) \end{cases};$$

$$r^{\text{BA}}_{\mathbf{P}}(i) = \begin{cases} \max\limits_{\mathcal{T} \in \text{MAST}(\mathbf{P}): \mathcal{T} \text{ contains } i} \Delta^{\mathbf{P}}_{m^*(\mathcal{T}),i} & \text{if } i \notin \text{BA}(\mathbf{P}) \\ 0 & \text{if } i \in \text{BA}(\mathbf{P}). \end{cases}$$

In the special case when $\mathbf{P}$ admits a Condorcet winner i^*, the three individual regrets above all reduce to the Condorcet individual regret, $r^{\text{CW}}_{\mathbf{P}}(i) = \Delta^{\mathbf{P}}_{i^*,i}$. In each case above, the cumulative regret of an algorithm $\mathcal{A}$ over T trials will then be given by

$$\mathcal{R}^{\mathsf{S}}_T(\mathcal{A}) \;=\; \textstyle\sum_{t=1}^{T} r^{\mathsf{S}}_{\mathbf{P}}(i_t, j_t)\,,$$

where the pairwise regret $r^{\mathsf{S}}_{\mathbf{P}}(i,j)$ can be the average regret $\frac{1}{2}\big(r^{\mathsf{S}}_{\mathbf{P}}(i) + r^{\mathsf{S}}_{\mathbf{P}}(j)\big)$, the strong regret $\max\big(r^{\mathsf{S}}_{\mathbf{P}}(i), r^{\mathsf{S}}_{\mathbf{P}}(j)\big)$, or the weak regret $\min\big(r^{\mathsf{S}}_{\mathbf{P}}(i), r^{\mathsf{S}}_{\mathbf{P}}(j)\big)$. Our regret bounds will hold for each of these forms of pairwise regret. In fact, our regret bounds hold for any measure of pairwise regret $r^{\mathsf{S}}_{\mathbf{P}}(i,j)$ that satisfies the following three conditions:

(i) $r^{\mathsf{S}}_{\mathbf{P}}(\cdot,\cdot)$ is *normalized*: $r^{\mathsf{S}}_{\mathbf{P}}(i,j) \in [0,1]\ \forall i,j$;

(ii) $r^{\mathsf{S}}_{\mathbf{P}}(\cdot,\cdot)$ is *symmetric*: $r^{\mathsf{S}}_{\mathbf{P}}(i,j) = r^{\mathsf{S}}_{\mathbf{P}}(j,i)\ \forall i,j$; and

(iii) $r^{\mathsf{S}}_{\mathbf{P}}(\cdot,\cdot)$ is *proper w.r.t.* S: $i,j \in \mathsf{S}(\mathbf{P}) \implies r^{\mathsf{S}}_{\mathbf{P}}(i,j) = 0$.

It is easy to verify that for the three tournament solutions above, the average, strong and weak pairwise regrets above all satisfy these conditions.[5,6]

3 UCB-TS: Generic Dueling Bandit Algorithm for Tournament Solutions

Algorithm. In Algorithm 1 we outline a generic dueling bandit algorithm, which we call UCB-TS, for identifying winners from a general tournament solution. The algorithm can be instantiated to specific tournament solutions by designing suitable selection procedures SELECTPROC-TS (more details below). The algorithm maintains a matrix $\mathbf{U}^t \in \mathbb{R}_+^{K\times K}$ of upper confidence bounds (UCBs) U^t_{ij} on the unknown pairwise preference probabilities P_{ij}. The UCBs are constructed by adding a confidence term to the current empirical estimate of P_{ij}; the exploration parameter $\alpha > \frac{1}{2}$ controls the exploration rate of the algorithm via the size of the confidence terms used. On each trial t, the algorithm selects a pair of arms (i_t, j_t) based on the current UCB matrix $\mathbf{U}^t$ using the selection procedure SELECTPROC-TS; on observing the preference feedback y_t, the algorithm then updates the UCBs for all pairs of arms (i,j) (the UCBs of all pairs (i,j) grow slowly with t so that pairs that have not been selected for a while have an increasing chance of being explored).

In order to instantiate the UCB-TS algorithm to a particular tournament solution S, the critical step is in designing the selection procedure SELECTPROC-TS in a manner that yields good regret bounds for a suitable regret measure w.r.t. S. Below we identify general conditions on SELECTPROC-TS that allow for $O(\ln T)$ anytime regret bounds to be obtained (we will design procedures satisfying these conditions for the three tournament solutions of interest in Section 4).

[4] The notion of regret used in [5] was slightly different.

[5] It is also easy to verify that defining the individual regrets as the minimum or average margin relative to all relevant arms in the tournament solution of interest (instead of the maximum margin as done above) also preserves these properties, and therefore our regret bounds hold for the resulting variants of regret as well.

[6] One can also consider defining the individual regrets simply in terms of *mistakes* relative to the target tournament solution of interest, e.g. $r^{\text{TC}}_{\mathbf{P}}(i) = \mathbf{1}(i \notin \text{TC}(\mathbf{P}))$, and define average/strong/weak pairwise regrets in terms of these; our bounds also apply in this case.

Algorithm 1 UCB-TS

Require: Selection procedure SELECTPROC-TS

Parameter: Exploration parameter $\alpha > \frac{1}{2}$

Initialize: $\forall\, (i,j) \in [K] \times [K]$:

$N^1_{ij} = 0$ // # times (i,j) has been compared; $W^1_{ij} = 0$ // # times i has won against j;

$$U^1_{ij} = \begin{cases} \frac{1}{2} & \text{if } i = j \\ 1 & \text{otherwise} \end{cases} \quad \text{// UCB for } P_{ij}.$$

For $t = 1, 2, \ldots$ **do:**

- Select $(i_t, j_t) \leftarrow$ SELECTPROC-TS$(\mathbf{U}^t)$
- Receive preference feedback $y_t \in \{0,1\}$
- Update counts: $\forall\, (i,j) \in [K] \times [K]$:

$$N^{t+1}_{ij} = \begin{cases} N^t_{ij} + 1 & \text{if } \{i,j\} = \{i_t, j_t\} \\ N^t_{ij} & \text{otherwise} \end{cases}; \quad W^{t+1}_{ij} = \begin{cases} W^t_{ij} + y_t & \text{if } (i,j) = (i_t, j_t) \\ W^t_{ij} + (1 - y_t) & \text{if } (i,j) = (j_t, i_t) \\ W^t_{ij} & \text{otherwise.} \end{cases}$$

- Update UCBs: $\forall\, (i,j) \in [K] \times [K]$:

$$U^{t+1}_{ij} = \begin{cases} \frac{1}{2} & \text{if } i = j \\ 1 & \text{if } i \neq j \text{ and } N^{t+1}_{ij} = 0 \\ \frac{W^{t+1}_{ij}}{N^{t+1}_{ij}} + \sqrt{\frac{\alpha \ln t}{N^{t+1}_{ij}}} & \text{otherwise.} \end{cases}$$

Regret Analysis. We show here that if the selection procedure SELECTPROC-TS satisfies two natural conditions w.r.t. a tournament solution S, namely the *safe identical-arms condition w.r.t.* S and the *safe distinct-arms condition w.r.t.* S, then the resulting instantiation of the UCB-TS algorithm has an $O(\ln T)$ regret bound for any regret measure that is normalized, symmetric, and proper w.r.t S. The first condition ensures that if the UCB matrix $\mathbf{U}$ given as input to SELECTPROC-TS in fact forms an element-wise upper bound on the true preference matrix $\mathbf{P}$ and SELECTPROC-TS returns two identical arms (i,i), then i must be in the winning set $\mathsf{S}(\mathbf{P})$. The second condition ensures that if $\mathbf{U}$ upper bounds $\mathbf{P}$ and SELECTPROC-TS returns two distinct arms (i,j), $i \neq j$, then either both i,j are in the winning set $\mathsf{S}(\mathbf{P})$, or the UCBs U_{ij}, U_{ji} are still loose (and (i,j) should be explored further).

Definition 6 (**Safe identical-arms condition**). *Let* $\mathsf{S} : \mathcal{P}_K \to 2^{[K]}$ *be a tournament solution. We will say a selection procedure* SELECTPROC-TS : $\mathbb{R}_+^{K\times K} \to [K] \times [K]$ *satisfies the* safe identical-arms condition w.r.t. S *if for all* $\mathbf{P} \in \mathcal{P}_K$, $\mathbf{U} \in \mathbb{R}_+^{K\times K}$ *such that* $P_{ij} \le U_{ij}\ \forall i,j$, *we have*

$$\text{SELECTPROC-TS}(\mathbf{U}) = (i,i) \implies i \in \mathsf{S}(\mathbf{P})\,.$$

Definition 7 (**Safe distinct-arms condition**). *Let* $\mathsf{S} : \mathcal{P}_K \to 2^{[K]}$ *be a tournament solution. We will say a selection procedure* SELECTPROC-TS : $\mathbb{R}_+^{K\times K} \to [K] \times [K]$ *satisfies the* safe distinct-arms condition w.r.t. S *if for all* $\mathbf{P} \in \mathcal{P}_K$, $\mathbf{U} \in \mathbb{R}_+^{K\times K}$ *such that* $P_{ij} \le U_{ij}\ \forall i,j$, *we have*

$$\text{SELECTPROC-TS}(\mathbf{U}) = (i,j),\ i \neq j \implies \big\{(i,j) \in \mathsf{S}(\mathbf{P}) \times \mathsf{S}(\mathbf{P})\big\}\ \text{or}\ \big\{U_{ij} + U_{ji} \ge 1 + \Delta^{\mathbf{P}}_{ij}\big\}\,.$$

In what follows, for $K \in \mathbb{Z}_+$, $\alpha > \frac{1}{2}$, and $\delta \in (0,1]$, we define

$$C(K,\alpha,\delta) = \Big(\frac{(4\alpha-1)K^2}{(2\alpha-1)\delta}\Big)^{1/(2\alpha-1)}\,.$$

This quantity, which also appears in the analysis of RUCB [6], acts as an initial time period beyond which all the UCBs U_{ij} upper bound P_{ij} w.h.p. We have the following result (proof in Appendix A):

Theorem 8 (**Regret bound for UCB-TS algorithm**). *Let* $\mathsf{S} : \mathcal{P}_K \to 2^{[K]}$ *be a tournament solution, and suppose the selection procedure* SELECTPROC-TS *used in the UCB-TS algorithm satisfies both the safe identical-arms condition w.r.t.* S *and the safe distinct-arms condition w.r.t.* S. *Let* $\mathbf{P} \in \mathcal{P}_K$, *and let* $r^{\mathsf{S}}_{\mathbf{P}}(i,j)$ *be any normalized, symmetric, proper regret measure w.r.t.* S. *Let* $\alpha > \frac{1}{2}$ *and* $\delta \in (0,1]$. *Then with probability at least* $1-\delta$ *(over the feedback* y_t *drawn randomly from* $\mathbf{P}$ *and any internal randomness in* SELECTPROC-TS*), the cumulative regret of the UCB-TS algorithm with exploration parameter* α *is upper bounded as*

$$\mathcal{R}^{\mathsf{S}}_T\big(\text{UCB-TS}(\alpha)\big) \quad \le \quad C(K,\alpha,\delta) + 4\alpha\,(\ln T)\Bigg(\sum_{i<j:(i,j)\notin \mathsf{S}(\mathbf{P})\times\mathsf{S}(\mathbf{P})} \frac{r^{\mathsf{S}}_{\mathbf{P}}(i,j)}{(\Delta^{\mathbf{P}}_{ij})^2}\Bigg)\,.$$

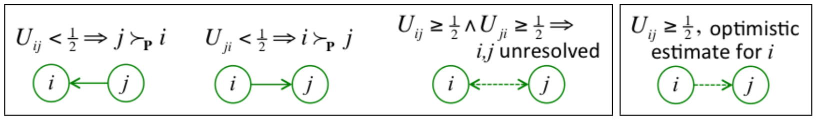


Figure 3: Inferences about the direction of preference between arms i and j under the true preference matrix $\mathbf{P}$ based on the UCBs U_{ij}, U_{ji}, assuming that P_{ij}, P_{ji} are upper bounded by U_{ij}, U_{ji}.

4 Dueling Bandit Algorithms for Top Cycle, Uncovered Set, and Banks Set

Below we give selection procedures satisfying both the safe identical-arms condition and the safe distinct-arms condition above w.r.t. the top cycle, uncovered set, and Banks set, which immediately yield dueling bandit algorithms with $O(\ln T)$ regret bounds w.r.t. these tournament solutions. An instantiation of our framework to the Copeland set is also discussed in Appendix E.

The selection procedure for each tournament solution is closely related to the corresponding winner determination algorithm for that tournament solution; however, while a standard winner determination algorithm would have access to the actual tournament $\mathcal{T}_{\mathbf{P}}$, the selection procedures we design can only guess (with high confidence) the preference directions between some pairs of arms based on the UCB matrix $\mathbf{U}$. In particular, if the entries of $\mathbf{U}$ actually upper bound those of $\mathbf{P}$, then for any pair of arms i and j, one of the following must be true (see also Figure 3):

- $U_{ij} < \frac{1}{2}$, in which case $P_{ij} \leq U_{ij} < \frac{1}{2}$ and therefore $j \succ_{\mathbf{P}} i$;
- $U_{ji} < \frac{1}{2}$, in which case $P_{ji} \leq U_{ji} < \frac{1}{2}$ and therefore $i \succ_{\mathbf{P}} j$;
- $U_{ij} \geq \frac{1}{2}$ and $U_{ji} \geq \frac{1}{2}$, in which case the direction of preference between i and j in $\mathcal{T}_{\mathbf{P}}$ is unresolved.

The selection procedures we design manage the exploration-exploitation tradeoff by adopting an *optimism followed by pessimism* approach, similar to that used in the design of the RUCB and CCB algorithms [6, 11]. Specifically, our selection procedures first optimistically identify a potential winning arm a based on the UCBs $\mathbf{U}$ (by optimistically setting directions of any unresolved edges in $\mathcal{T}_{\mathbf{P}}$ in favor of the arm being considered; see Figure 3). Once a putative winning arm a is identified, the selection procedures then pessimistically find an arm b that has the greatest chance of invalidating a as a winning arm, and select the pair (a, b) for comparison.

4.1 UCB-TC: Dueling Bandit Algorithm for Top Cycle

The selection procedure SELECTPROC-TC (Algorithm 2), when instantiated in the UCB-TS template, yields the UCB-TC dueling bandit algorithm. Intuitively, SELECTPROC-TC constructs an optimistic estimate A of the top cycle based on the UCBs $\mathbf{U}$ (line 2), and selects a potential winning arm a from A (line 3); if there is no unresolved arm against a (line 5), then it returns (a, a) for comparison, else it selects the best-performing unresolved opponent b (line 8) and returns (a, b) for comparison. We have the following result (see Appendix B for a proof):

Theorem 9 (SELECTPROC-TC satisfies safety conditions w.r.t. TC). SELECTPROC-TC *satisfies both the safe identical-arms condition and the safe distinct-arms condition w.r.t.* TC.

By virtue of Theorem 8, this immediately yields the following regret bound for UCB-TC:

Corollary 10 (Regret bound for UCB-TC algorithm). *Let $\mathbf{P} \in \mathcal{P}_K$. Let $\alpha > \frac{1}{2}$ and $\delta \in (0, 1]$. Then with probability at least $1 - \delta$, the cumulative regret of UCB-TC w.r.t. the top cycle satisfies*

$$\mathcal{R}_T^{\mathrm{TC}}\big(\text{UCB-TC}(\alpha)\big) \quad \leq \quad C(K, \alpha, \delta) + 4\alpha\,(\ln T) \left(\sum_{i<j:(i,j)\notin \mathrm{TC}(\mathbf{P})\times\mathrm{TC}(\mathbf{P})} \frac{r_{\mathbf{P}}^{\mathrm{TC}}(i,j)}{(\Delta_{ij}^{\mathbf{P}})^2} \right).$$

4.2 UCB-UC: Dueling Bandit Algorithm for Uncovered Set

The selection procedure SELECTPROC-UC (Algorithm 3), when instantiated in the UCB-TS template, yields the UCB-UC dueling bandit algorithm. SELECTPROC-UC relies on the property that an uncovered arm beats every other arm either directly or via an intermediary [12]. SELECTPROC-UC optimistically identifies such a potentially uncovered arm a based on the UCBs $\mathbf{U}$ (line 5); if it can be resolved that a is indeed uncovered (line 7), then it returns (a, a), else it selects the best-performing unresolved opponent b when available (line 11), or an arbitrary opponent b otherwise (line 13), and returns (a, b). We have the following result (see Appendix C for a proof):

Algorithm 2 SELECTPROC-TC

Input: UCB matrix $\mathbf{U} \in \mathbb{R}_+^{K \times K}$
Let $A \subseteq [K]$ be any minimal set satisfying $U_{ij} \geq \frac{1}{2}\ \forall i \in A, j \notin A$
Select any $a \in \mathrm{argmax}_{i \in A} \min_{j \notin A} U_{ij}$
$B \leftarrow \{i \neq a : U_{ai} \geq \frac{1}{2} \wedge U_{ia} \geq \frac{1}{2}\}$
if $B = \emptyset$ **then**
    Return (a, a)
else
    Select any $b \in \mathrm{argmax}_{i \in B} U_{ia}$
    Return (a, b)
end if

Algorithm 3 SELECTPROC-UC

Input: UCB matrix $\mathbf{U} \in \mathbb{R}_+^{K \times K}$
for $i = 1$ **to** K **do**
    $y(i) \leftarrow \sum_j \mathbf{1}(U_{ij} \geq \frac{1}{2}) + \sum_{j,k} \mathbf{1}(U_{ij} \geq \frac{1}{2} \wedge U_{jk} \geq \frac{1}{2})$
end for
Select any $a \in \mathrm{argmax}_i\, y(i)$
$B \leftarrow \{i \neq a : U_{ai} \geq \frac{1}{2} \wedge U_{ia} \geq \frac{1}{2}\}$
if $\big(\forall i \neq a : (U_{ia} < \frac{1}{2}) \vee (\exists j : U_{ij} < \frac{1}{2} \wedge U_{ja} < \frac{1}{2})\big)$ **then**
    Return (a, a)
else
    if $B \neq \emptyset$ **then**
        Select any $b \in \mathrm{argmax}_{i \in B} U_{ia}$
    else
        Select any $b \neq a$
    end if
    Return (a, b)
end if

Algorithm 4 SELECTPROC-BA

1: **Input:** UCB matrix $\mathbf{U} \in \mathbb{R}_+^{K \times K}$
2: Select any $j_1 \in [K]$
3: $\mathcal{J} \leftarrow \{j_1\}$ // Initialize candidate Banks trajectory
4: $s \leftarrow 1$ // Initialize size of candidate Banks trajectory
5: traj_found = FALSE
6: **while** NOT(traj_found) **do**
7:     $C \leftarrow \{i \notin \mathcal{J} : U_{ij} > \frac{1}{2}\ \ \forall j \in \mathcal{J}\}$
8:     **if** $C = \emptyset$ **then**
9:         traj_found = TRUE
10:         **break**
11:     **else**
12:         $j_{s+1} \in \mathrm{argmax}_{i \in C}(\min_{j \in \mathcal{J}} U_{ij})$
13:         $\mathcal{J} \leftarrow \mathcal{J} \cup \{j_{s+1}\}$
14:         $s \leftarrow s + 1$
15:     **end if**
16: **end while**
17: **if** $\big(\forall 1 \leq q < r \leq s$: $U_{j_q, j_r} < \frac{1}{2}\big)$ **then**
18:     $a \leftarrow j_s$
19:     Return (a, a)
20: **else**
21:     Select any $(\widetilde{q}, \widetilde{r}) \in \underset{(q,r): 1 \leq q < r \leq s}{\arg\max} U_{j_q, j_r}$
22:     $(a, b) \leftarrow (j_{\widetilde{q}}, j_{\widetilde{r}})$
23:     Return (a, b)
24: **end if**

Theorem 11 (SELECTPROC-UC satisfies safety conditions w.r.t. UC). SELECTPROC-UC *satisfies both the safe identical-arms condition and the safe distinct-arms condition w.r.t.* UC.

Again, by virtue of Theorem 8, this immediately yields the following regret bound for UCB-UC:

Corollary 12 (Regret bound for UCB-UC algorithm). *Let $\mathbf{P} \in \mathcal{P}_K$. Let $\alpha > \frac{1}{2}$ and $\delta \in (0, 1]$. Then with probability at least $1 - \delta$, the cumulative regret of UCB-UC w.r.t. the uncovered set satisfies*

$$\mathcal{R}_T^{\mathrm{UC}}\big(\text{UCB-UC}(\alpha)\big) \quad \leq \quad C(K, \alpha, \delta) + 4\alpha\,(\ln T)\left(\sum_{i<j:(i,j) \notin \mathrm{UC}(\mathbf{P}) \times \mathrm{UC}(\mathbf{P})} \frac{r_{\mathbf{P}}^{\mathrm{UC}}(i,j)}{(\Delta_{ij}^{\mathbf{P}})^2}\right).$$

4.3 UCB-BA: Dueling Bandit Algorithm for Banks Set

The selection procedure SELECTPROC-BA (Algorithm 4), when instantiated in the UCB-TS template, yields the UCB-BA dueling bandit algorithm. Intuitively, SELECTPROC-BA first constructs an optimistic candidate maximal acyclic subtournament (set $\mathcal{J}$; also called a Banks trajectory) based on the UCBs $\mathbf{U}$ (lines 2–16). If this subtournament is completely resolved (line 17), then its maximal arm a is picked and (a, a) is returned; if not, an unresolved pair (a, b) is returned that is most likely to fail the acyclicity/transitivity property. We have the following result (see Appendix D for a proof):

Theorem 13 (SELECTPROC-BA satisfies safety conditions w.r.t. BA). SELECTPROC-BA *satisfies both the safe identical-arms condition and the safe distinct-arms condition w.r.t.* BA.

Again, by virtue of Theorem 8, this immediately yields the following regret bound for UCB-BA:

Corollary 14 (Regret bound for UCB-BA algorithm). *Let $\mathbf{P} \in \mathcal{P}_K$. Let $\alpha > \frac{1}{2}$ and $\delta \in (0, 1]$. Then with probability at least $1 - \delta$, the cumulative regret of UCB-BA w.r.t. the Banks set satisfies*

$$\mathcal{R}_T^{\mathrm{BA}}\big(\text{UCB-BA}(\alpha)\big) \quad \leq \quad C(K, \alpha, \delta) + 4\alpha\,(\ln T)\left(\sum_{i<j:(i,j) \notin \mathrm{BA}(\mathbf{P}) \times \mathrm{BA}(\mathbf{P})} \frac{r_{\mathbf{P}}^{\mathrm{BA}}(i,j)}{(\Delta_{ij}^{\mathbf{P}})^2}\right).$$

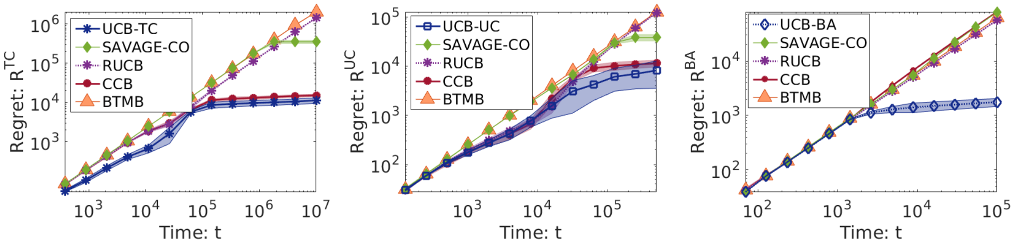


Figure 4: Regret performance of our algorithms compared to BTMB, RUCB, SAVAGE-CO, and CCB. Results are averaged over 10 independent runs; light colored bands represent one standard error. **Left:** Top cycle regret of UCB-TC on $\mathbf{P}^{\text{MSLR}}$. **Middle:** Uncovered set regret of UCB-UC on $\mathbf{P}^{\text{Tennis}}$. **Right:** Banks set regret of UCB-BA on $\mathbf{P}^{\text{Hudry}}$. See Appendix F.2 for additional results.

5 Experiments

Here we provide an empirical evaluation of the performance of the proposed dueling bandit algorithms. We used the following three preference matrices in our experiments, one of which is synthetic and two real-world, and none of which posesses a Condorcet winner:

- $\mathbf{P}^{\text{Hudry}} \in \mathcal{P}_{13}$: This is constructed from the Hudry tournament shown in Figure 2(b); as noted earlier, this is the smallest tournament whose Copeland set and Banks set are disjoint [14]. Details of this preference matrix can be found in Appendix F.1.1.
- $\mathbf{P}^{\text{Tennis}} \in \mathcal{P}_{8}$: This is constructed from real data collected from the Association of Tennis Professionals' (ATP's) website on outcomes of tennis matches played among 8 well-known professional tennis players. The tournament associated with $\mathbf{P}^{\text{Tennis}}$ is shown in Figure 2(c); further details of this preference matrix can be found in Appendix F.1.2.
- $\mathbf{P}^{\text{MSLR}} \in \mathcal{P}_{16}$: This is constructed from real data from the Microsoft Learning to Rank (MSLR) Web10K data set. Further details can be found in Appendix F.1.3.

We compared the performance of our algorithms, UCB-TC, UCB-BA, and UCB-UC, with four previous dueling bandit algorithms: BTMB [2], RUCB [6], SAVAGE-CO [3], and CCB [11].[7] In each case, we assessed the algorithms in terms of average pairwise regret relative to the target tournament solution of interest (see Section 2), averaged over 10 independent runs. A sample of the results is shown in Figure 4; as can be seen, the proposed algorithms UCB-TC, UCB-UC, and UCB-BA generally outperform existing baselines in terms of minimizing regret relative to the top cycle, the uncovered set, and the Banks set, respectively. Additional results, including results with the Copeland set variant of our algorithm, UCB-CO, can be found in Appendix F.2.

6 Conclusion

In this paper, we have proposed the use of general tournament solutions as sets of 'winning' arms in stochastic dueling bandit problems, with the advantage that these tournament solutions always exist and can be used to define winners according to criteria that are most relevant to a given dueling bandit setting. We have developed a UCB-style family of algorithms for such general tournament solutions, and have shown $O(\ln T)$ anytime regret bounds for the algorithm instantiated to the top cycle, uncovered set, and Banks set (as well as the Copeland set; see Appendix E).

While our approach has an appealing modular structure both algorithmically and in our proofs, an open question concerns the optimality of our regret bounds in their dependence on the number of arms K. For the Condorcet winner, the MergeRUCB algorithm [7] has an anytime regret bound of the form $O(K \ln T)$; for the Copeland set, the SCB algorithm [11] has an anytime regret bound of the form $O(K \ln K \ln T)$. In the worst case, our regret bounds are of the form $O(K^2 \ln T)$. Is it possible that for the top cycle, uncovered set, and Banks set, one can also show an $\Omega(K^2 \ln T)$ lower bound on the regret? Or can our regret bounds or algorithms be improved? We leave a detailed investigation of this issue to future work.

Acknowledgments. Thanks to the anonymous reviewers for helpful comments and suggestions. SR thanks Google for a travel grant to present this work at the conference.

[7]For all the UCB-based algorithms (including our algorithms, RUCB, and CCB), we set the exploration parameter α to 0.51; for SAVAGE-CO, we set the confidence parameter δ to $1/T$; and for BTMB, we set δ to $1/T$ and chose γ to satisfy the γ-relaxed stochastic transitivity property for each preference matrix.

References

[1] Yisong Yue, Josef Broder, Robert Kleinberg, and Thorsten Joachims. The K-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78(5):1538–1556, 2012.

[2] Yisong Yue and Thorsten Joachims. Beat the mean bandit. In *Proceedings of the 28th International Conference on Machine Learning*, 2011.

[3] Tanguy Urvoy, Fabrice Clerot, Raphael Féraud, and Sami Naamane. Generic exploration and k-armed voting bandits. In *Proceedings of the 30th International Conference on Machine Learning*, 2013.

[4] Róbert Busa-Fekete, Balazs Szorenyi, Weiwei Cheng, Paul Weng, and Eyke Hüllermeier. Top-k selection based on adaptive sampling of noisy preferences. In *Proceedings of the 30th International Conference on Machine Learning*, 2013.

[5] Nir Ailon, Zohar Karnin, and Thorsten Joachims. Reducing dueling bandits to cardinal bandits. In *Proceedings of the 31st International Conference on Machine Learning*, 2014.

[6] Masrour Zoghi, Shimon Whiteson, Remi Munos, and Maarten de Rijke. Relative upper confidence bound for the k-armed dueling bandit problem. In *Proceedings of the 31st International Conference on Machine Learning*, 2014.

[7] Masrour Zoghi, Shimon Whiteson, and Maarten de Rijke. MergeRUCB: A method for large-scale online ranker evaluation. In *Proceedings of the 8th ACM International Conference on Web Search and Data Mining*, 2015.

[8] Kevin Jamieson, Sumeet Katariya, Atul Deshpande, and Robert Nowak. Sparse dueling bandits. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, 2015.

[9] Junpei Komiyama, Junya Honda, Hisashi Kashima, and Hiroshi Nakagawa. Regret lower bound and optimal algorithm in dueling bandit problem. In *Proceedings of the 28th Conference on Learning Theory*, 2015.

[10] Miroslav Dudık, Katja Hofmann, Robert E Schapire, Aleksandrs Slivkins, and Masrour Zoghi. Contextual dueling bandits. In *Proceedings of the 28th Conference on Learning Theory*, 2015.

[11] Masrour Zoghi, Zohar S. Karnin, Shimon Whiteson, and Maarten de Rijke. Copeland dueling bandits. In *Advances in Neural Information Processing Systems 28*, 2015.

[12] Felix Brandt, Markus Brill, and Paul Harrenstein. Tournament solutions. In *Handbook of Computational Social Choice*. Cambridge University Press, 2016.

[13] Felix Brandt, Andre Dau, and Hans Georg Seedig. Bounds on the disparity and separation of tournament solutions. *Discrete Applied Mathematics*, 187:41–49, 2015.

[14] Olivier Hudry. A smallest tournament for which the Banks set and the Copeland set are disjoint. *Social Choice and Welfare*, 16(1):137–143, 1999.

[15] Kenneth A Shepsle and Barry R Weingast. Uncovered sets and sophisticated voting outcomes with implications for agenda institutions. *American Journal of Political Science*, pages 49–74, 1984.

[16] Kevin Jamieson and Robert Nowak. Active ranking using pairwise comparisons. In *Advances in Neural Information Processing Systems*, 2011.

Dueling Bandits: Beyond Condorcet Winners to General Tournament Solutions

A Proof of Theorem 8

We will make use of the following two lemmas. The first lemma upper bounds the UCBs of pairs of arms that have been played sufficiently many times; the second lemma, adapted from [6], states that with high probability, after an initial transient period of play, all the UCBs upper bound the actual pairwise probabilities.

Lemma 15. *Let $(i, j) \in [K] \times [K]$ and $\Delta > 0$. If at any iteration t of the UCB-TS algorithm, run with any selection procedure* SELECTPROC-TS*, we have $N^t_{ij} > \frac{4\alpha \ln(t)}{\Delta^2}$, then*

$$U^t_{ij} + U^t_{ji} < 1 + \Delta\,.$$

PROOF. Let $N^t_{ij} > \frac{4\alpha \ln(t)}{\Delta^2}$. Then we have

$$\begin{aligned} U^t_{ij} + U^t_{ji} &= \frac{W^t_{ij}}{N^t_{ij}} + \sqrt{\frac{\alpha \ln(t)}{N^t_{ij}}} + \frac{W^t_{ji}}{N^t_{ji}} + \sqrt{\frac{\alpha \ln(t)}{N^t_{ji}}}\,, \quad \text{by definition of the UCBs} \\ &= 1 + 2\sqrt{\frac{\alpha \ln(t)}{N^t_{ij}}}\,, \quad \text{since } W^t_{ij} + W^t_{ji} \;=\; N^t_{ij} \;=\; N^t_{ji} \\ &< 1 + \Delta\,, \quad \text{by assumption on } N^t_{ij}. \end{aligned}$$

□

Lemma 16 (Adapted from Zoghi et al., 2014 [6])**.** *Let $\mathbf{P} \in \mathcal{P}_K$, $\alpha > \frac{1}{2}$, and $\delta \in (0, \frac{1}{2}]$, and let $C(K, \alpha, \delta) = \big(\frac{(4\alpha-1)K^2}{(2\alpha-1)\delta}\big)^{1/(2\alpha-1)}$. Then for any selection procedure* SELECTPROC-TS*, with probability at least $1 - \delta$ (over the feedback y_t drawn from $\mathbf{P}$ and any internal randomness in* SELECTPROC-TS*), the UCBs constructed by the UCB-TS algorithm satisfy*

$$\forall t \geq C(K, \alpha, \delta): \quad P_{ij} \leq U^t_{ij} \;\forall (i, j) \in [K] \times [K]\,.$$

The proof of the above lemma follows that of Zoghi et al. [6].

We are now ready to prove Theorem 8.

PROOF. [Proof of Theorem 8]
We will bound the regret conditioned on the 'good' event $\mathcal{E}$ that

$$\forall t \geq C(K, \alpha, \delta): \quad P_{ij} \leq U^t_{ij} \;\forall (i, j) \in [K] \times [K]\,;$$

the result will then follow from Lemma 16. In the following, let

$$T_0 = \lfloor C(K, \alpha, \delta) \rfloor\,.$$

As in the description of the UCB-TS algorithm, for each $\tau \in \mathbb{Z}_+$ and $(i, j) \in [K] \times [K]$, let N^τ_{ij} denote the number of trials up to trial τ in which the pair (i, j) is compared:

$$N^\tau_{ij} = \sum_{t=1}^{\tau} \Big(\mathbf{1}\big((i_t, j_t) = (i, j)\big) + \mathbf{1}\big((i_t, j_t) = (j, i)\big) \Big)\,.$$

Then we can write the regret as

$$\begin{aligned}
\mathcal{R}_T^{\mathsf{S}}(\text{UCB-TS}(\alpha)) &= \sum_{t=1}^{T} r_{\mathbf{P}}^{\mathsf{S}}(i_t, j_t) \\
&= \sum_{i,j} \mathbf{1}\big((i_t, j_t) = (i,j)\big) \cdot r_{\mathbf{P}}^{\mathsf{S}}(i,j) \\
&= \sum_{i \le j} N_{ij}^T \, r_{\mathbf{P}}^{\mathsf{S}}(i,j)\,, \qquad \text{since } r_{\mathbf{P}}^{\mathsf{S}}(\cdot,\cdot) \text{ is symmetric} \\
&= \sum_{i \le j : (i,j) \notin \mathsf{S}(\mathbf{P}) \times \mathsf{S}(\mathbf{P})} N_{ij}^T \, r_{\mathbf{P}}^{\mathsf{S}}(i,j)\,, \qquad \text{since } r_{\mathbf{P}}^{\mathsf{S}}(\cdot,\cdot) \text{ is proper w.r.t. } \mathsf{S} \\
&= \sum_{i \notin \mathsf{S}(\mathbf{P})} N_{ii}^T \, r_{\mathbf{P}}^{\mathsf{S}}(i,i) + \sum_{i < j : (i,j) \notin \mathsf{S}(\mathbf{P}) \times \mathsf{S}(\mathbf{P})} N_{ij}^T \, r_{\mathbf{P}}^{\mathsf{S}}(i,j)\,. \qquad (1)
\end{aligned}$$

We will show that, conditioned on the event $\mathcal{E}$, each of the above two terms can be bounded.

First, consider any arm $i \notin \mathsf{S}(\mathbf{P})$. Then, conditioned on $\mathcal{E}$, the safe identical-arms property of the selection procedure and Lemma 16 together ensure that after T_0 trials, arm i is not selected for comparison with itself. Therefore, conditioned on $\mathcal{E}$, we have

$$N_{ii}^T = N_{ii}^{T_0}\,. \qquad (2)$$

Next, consider any pair of arms $i < j$ with $(i,j) \notin \mathsf{S}(\mathbf{P}) \times \mathsf{S}(\mathbf{P})$. In this case, conditioned on $\mathcal{E}$, the safe distinct-arms property of the selection procedure and Lemmas 15-16 together ensure that after T_0 trials, whenever arms i and j are compared on some trial t, we must have $N_{ij}^t \le \frac{4\alpha \ln(t)}{(\Delta_{ij}^{\mathbf{P}})^2}$. Therefore, defining

$$\begin{aligned}
A_{ij} &= \Big\{ T_0 < t \le T : \; N_{ij}^t \le \tfrac{4\alpha \ln(t)}{(\Delta_{ij}^{\mathbf{P}})^2} \Big\} \\
T_{ij} &= \max\big\{ t : t \in A_{ij} \big\}\,,
\end{aligned}$$

we have that, conditioned on $\mathcal{E}$,

$$\begin{aligned}
N_{ij}^T &= N_{ij}^{T_0} + \sum_{t \in A_{ij}} \Big(\mathbf{1}\big((i_t,j_t) = (i,j)\big) + \mathbf{1}\big((i_t,j_t) = (j,i)\big) \Big) \\
&\le N_{ij}^{T_0} + \sum_{t=1}^{T_{ij}} \Big(\mathbf{1}\big((i_t,j_t) = (i,j)\big) + \mathbf{1}\big((i_t,j_t) = (j,i)\big) \Big)\,, \qquad \text{by definition of } T_{ij} \\
&= N_{ij}^{T_0} + N_{ij}^{T_{ij}} \\
&\le N_{ij}^{T_0} + \frac{4\alpha \ln(T_{ij})}{(\Delta_{ij}^{\mathbf{P}})^2}\,, \qquad \text{since } T_{ij} \in A_{ij} \\
&\le N_{ij}^{T_0} + \frac{4\alpha \ln(T)}{(\Delta_{ij}^{\mathbf{P}})^2}\,, \qquad \text{since } T_{ij} \le T. \qquad (3)
\end{aligned}$$

Thus, combining Eqs. (1-3), we have that, conditioned on $\mathcal{E}$,

$$\begin{aligned}
\mathcal{R}_T^{\mathsf{S}}(\text{UCB-TS}(\alpha)) &\le \sum_{i \notin \mathsf{S}(\mathbf{P})} N_{ii}^{T_0} \, r_{\mathbf{P}}^{\mathsf{S}}(i,i) + \sum_{i < j : (i,j) \notin \mathsf{S}(\mathbf{P}) \times \mathsf{S}(\mathbf{P})} \Big(N_{ij}^{T_0} + \frac{4\alpha \ln(T)}{(\Delta_{ij}^{\mathbf{P}})^2} \Big) r_{\mathbf{P}}^{\mathsf{S}}(i,j) \\
&\le \sum_{i \le j : (i,j) \notin \mathsf{S}(\mathbf{P}) \times \mathsf{S}(\mathbf{P})} N_{ij}^{T_0} + \sum_{i < j : (i,j) \notin \mathsf{S}(\mathbf{P}) \times \mathsf{S}(\mathbf{P})} \Big(\frac{4\alpha \ln(T)}{(\Delta_{ij}^{\mathbf{P}})^2} \Big) r_{\mathbf{P}}^{\mathsf{S}}(i,j)\,, \\
&\qquad \qquad \text{since } r_{\mathbf{P}}^{\mathsf{S}}(\cdot,\cdot) \text{ is normalized} \\
&= T_0 + \sum_{i < j : (i,j) \notin \mathsf{S}(\mathbf{P}) \times \mathsf{S}(\mathbf{P})} \Big(\frac{4\alpha \ln(T)}{(\Delta_{ij}^{\mathbf{P}})^2} \Big) r_{\mathbf{P}}^{\mathsf{S}}(i,j)\,.
\end{aligned}$$

The result follows. □

B Proof of Theorem 9

PROOF. We first prove that SELECTPROC-TC satisfies the safe identical-arms condition w.r.t. TC. Let $\mathbf{P} \in \mathcal{P}_K, \mathbf{U} \in \mathbb{R}_+^{K\times K}$ be such that $P_{ij} \leq U_{ij}\ \forall i,j$, and suppose that SELECTPROC-TC$(\mathbf{U}) = (a,a)$. We will show that $a \in \text{TC}(\mathbf{P})$. Let the sets A, B be defined as in SELECTPROC-TC. The identical-arms pair (a,a) must be returned via line 6 of the procedure, and therefore the condition in line 5 must be satisfied, i.e. the set B must be empty. Let, if possible, $a \notin \text{TC}(\mathbf{P})$ (we will show this leads to a contradiction). Then for any arm c, we have

$$\begin{aligned} c \in \text{TC}(\mathbf{P}) &\implies P_{ca} > \tfrac{1}{2}\,, \quad \text{by our assumption that } a \notin \text{TC}(\mathbf{P}) \\ &\implies U_{ca} > \tfrac{1}{2}\,, \quad \text{since } U_{ca} \geq P_{ca} \\ &\implies U_{ac} < \tfrac{1}{2}\,, \quad \text{since } B = \emptyset \text{ and therefore } c \notin B \\ &\implies c \in A\,, \quad \text{since } U_{ak} \geq \tfrac{1}{2}\ \forall k \notin A. \end{aligned}$$

Thus $\text{TC}(\mathbf{P}) \subseteq A$; in fact, since by our assumption, $a \in A \setminus \text{TC}(\mathbf{P})$, we have strict containment: $\text{TC}(\mathbf{P}) \subsetneq A$. Moreover, we have that for all $i \in \text{TC}(\mathbf{P}), j \notin \text{TC}(\mathbf{P})$, $U_{ij} \geq P_{ij} > \frac{1}{2}$. Thus the set $A' = \text{TC}(\mathbf{P})$ contradicts the minimality property in the definition of A, and therefore our assumption that $a \notin \text{TC}(\mathbf{P})$ must be false, i.e. we must have $a \in \text{TC}(\mathbf{P})$. This establishes that SELECTPROC-TC satisfies the safe identical-arms condition w.r.t. TC.

Next, we prove that SELECTPROC-TC satisfies the safe distinct-arms condition w.r.t. TC. Again, let $\mathbf{P} \in \mathcal{P}_K, \mathbf{U} \in \mathbb{R}_+^{K\times K}$ be such that $P_{ij} \leq U_{ij}\ \forall i,j$, and now suppose that SELECTPROC-TC$(\mathbf{U}) = (a,b)$ with $a \neq b$. Let the set B be defined as in SELECTPROC-TC. Now, (a,b) must be returned via line 9, and therefore we must have $b \in B$. By definition of B, this implies that we have both $U_{ab} \geq \frac{1}{2}$ and $U_{ba} \geq \frac{1}{2}$. Thus we have

$$\begin{aligned} U_{ab} + U_{ba} &= \max(U_{ab}, U_{ba}) + \min(U_{ab}, U_{ba}) \\ &\geq \max(P_{ab}, P_{ba}) + \tfrac{1}{2} \\ &= (\tfrac{1}{2} + \Delta_{ab}^{\mathbf{P}}) + \tfrac{1}{2} \\ &= 1 + \Delta_{ab}^{\mathbf{P}}\,. \end{aligned}$$

This establishes that SELECTPROC-TC satisfies the safe distinct-arms condition w.r.t. TC (in fact, w.r.t. any tournament solution). □

C Proof of Theorem 11

We will need the following lemma:

Lemma 17. *Let $\mathbf{U} \in \mathbb{R}_+^{K\times K}$ be such that $U_{ij} + U_{ji} \geq 1\ \forall i,j$. Let the set B be constructed from $\mathbf{U}$ as in line 6 of* SELECTPROC-UC. *If the condition in line 7 of* SELECTPROC-UC *is not satisfied, then $B \neq \emptyset$.*

PROOF. As in SELECTPROC-UC, define

$$\begin{aligned} y(i) &= \sum_{j=1}^{K} \mathbf{1}(U_{ij} \geq \tfrac{1}{2}) + \sum_{i=1}^{K}\sum_{k=1}^{K} \mathbf{1}(U_{ij} \geq \tfrac{1}{2} \wedge U_{jk} \geq \tfrac{1}{2}) \quad \forall i \in [K] \\ a &\in \operatorname{argmax}_i y(i)\,. \end{aligned}$$

Suppose the condition in line 7 of SELECTPROC-UC is not satisfied, i.e. suppose $\exists i \neq a$ such that $U_{ia} \geq \frac{1}{2}$, and for all j, either $U_{ij} \geq \frac{1}{2}$ or $U_{ja} \geq \frac{1}{2}$. We will show that the set B, defined as

$$B = \left\{ i \neq a : U_{ai} \geq \tfrac{1}{2} \wedge U_{ia} \geq \tfrac{1}{2} \right\},$$

is non-empty.

Let, if possible, B be empty. Then for all arms $i \neq a$, either $U_{ai} < \frac{1}{2}$ and $U_{ia} \geq \frac{1}{2}$, or $U_{ai} \geq \frac{1}{2}$ and $U_{ia} < \frac{1}{2}$ (note that since $U_{ai} + U_{ia} \geq 1$, we cannot have $U_{ai} < \frac{1}{2}$ and $U_{ia} < \frac{1}{2}$). Thus all arms $i \neq a$ are 'resolved' against a (under $\mathbf{U}$), and can be partitioned into a set C of arms that 'beat' a (under $\mathbf{U}$), and a set D of arms that 'lose' to a (under $\mathbf{U}$):

$$\begin{aligned} C &= \left\{ i \neq a : U_{ai} < \tfrac{1}{2} \wedge U_{ia} \geq \tfrac{1}{2} \right\} \\ D &= \left\{ i \neq a : U_{ai} \geq \tfrac{1}{2} \wedge U_{ia} < \tfrac{1}{2} \right\}. \end{aligned}$$

Next, we claim that there is an arm $c \in C$ s.t. $U_{cd} \geq \frac{1}{2} \; \forall d \in D$. Indeed, suppose not; then we must have $\forall c' \in C, \exists d' \in D$ s.t. $U_{c'd'} < \frac{1}{2}$. This means that a beats all arms in D (under $\mathbf{U}$) directly (by definition of D), and beats all arms in C (under $\mathbf{U}$) via an intermediary in D. But this contradicts the assumption that the condition in line 7 is not satisfied. Therefore, there must be an arm $c \in C$ s.t. $U_{cd} \geq \frac{1}{2} \; \forall d \in D$.

Now, consider $y(a)$ and $y(c)$:

$$
\begin{aligned}
y(a) &= \sum_{i \in [K]} \mathbf{1}(U_{ai} \geq \tfrac{1}{2}) + \sum_{j \in [K],\, k \in [K]} \mathbf{1}(U_{aj} \geq \tfrac{1}{2} \wedge U_{jk} \geq \tfrac{1}{2}) \\
&= \mathbf{1}(U_{aa} \geq \tfrac{1}{2}) + \sum_{i \in D} \mathbf{1}(U_{ai} \geq \tfrac{1}{2}) + \mathbf{1}(U_{aa} \geq \tfrac{1}{2} \wedge U_{aa} \geq \tfrac{1}{2}) + \sum_{j \in D,\, k \in C \cup D} \mathbf{1}(U_{aj} \geq \tfrac{1}{2} \wedge U_{jk} \geq \tfrac{1}{2}) \\
&= 1 + |D| + 1 + \sum_{j \in D,\, k \in C \cup D} \mathbf{1}(U_{jk} \geq \tfrac{1}{2}) \\
&= 2 + |D| + \sum_{j \in D,\, k \in C \cup D} \mathbf{1}(U_{jk} \geq \tfrac{1}{2})\,; \\
y(c) &= \sum_{i \in [K]} \mathbf{1}(U_{ci} \geq \tfrac{1}{2}) + \sum_{j \in [K],\, k \in [K]} \mathbf{1}(U_{cj} \geq \tfrac{1}{2} \wedge U_{jk} \geq \tfrac{1}{2}) \\
&\geq \mathbf{1}(U_{cc} \geq \tfrac{1}{2}) + \mathbf{1}(U_{ca} \geq \tfrac{1}{2}) + \sum_{i \in D} \mathbf{1}(U_{ci} \geq \tfrac{1}{2}) + \mathbf{1}(U_{cc} \geq \tfrac{1}{2} \wedge U_{cc} \geq \tfrac{1}{2}) \\
&\quad + \mathbf{1}(U_{cc} \geq \tfrac{1}{2} \wedge U_{ca} \geq \tfrac{1}{2}) + \mathbf{1}(U_{ca} \geq \tfrac{1}{2} \wedge U_{aa} \geq \tfrac{1}{2}) + \sum_{j \in D,\, k \in C \cup D} \mathbf{1}(U_{cj} \geq \tfrac{1}{2} \wedge U_{jk} \geq \tfrac{1}{2}) \\
&= 1 + 1 + |D| + 1 + 1 + 1 + \sum_{j \in D,\, k \in C \cup D} \mathbf{1}(U_{jk} \geq \tfrac{1}{2}) \\
&= 5 + |D| + \sum_{j \in D,\, k \in C \cup D} \mathbf{1}(U_{jk} \geq \tfrac{1}{2})\,.
\end{aligned}
$$

This gives $y(c) > y(a)$. However this contradicts the choice of a in line 3. Therefore our assumption that B is empty must be false, i.e. it must be the case that $B \neq \emptyset$. □

We will also need the following characterization of uncovered arms, which says that an arm is uncovered if and only if it beats every other arm either directly or via an intermediary (see [12, 15]):

Lemma 18 (Shepsle and Weingast, 1984 [15]). *Let $\mathbf{P} \in \mathcal{P}_K$. Then $w \in \mathrm{UC}(\mathbf{P})$ if and only if for all $i \neq w$, either $w \succ_{\mathbf{P}} i$ or $\exists j \in [K]$ such that $w \succ_{\mathbf{P}} j$ and $j \succ_{\mathbf{P}} i$.*

We are now ready for the proof of Theorem 11.

PROOF. [Proof of Theorem 11]
We first prove that SELECTPROC-UC satisfies the safe identical-arms condition w.r.t. UC. Let $\mathbf{P} \in \mathcal{P}_K, \mathbf{U} \in \mathbb{R}_+^{K \times K}$ be such that $P_{ij} \leq U_{ij} \; \forall i, j$, and suppose that SELECTPROC-UC$(\mathbf{U}) = (a, a)$. Now, (a, a) must be returned via line 8 of the procedure, and therefore the condition in line 7 must be satisfied. In particular, this condition states that for all $i \neq a$, either $U_{ia} < \frac{1}{2}$, or $\exists j \in [K]$ such that both $U_{ja} < \frac{1}{2}$ and $U_{ij} < \frac{1}{2}$. This implies that for all $i \neq a$, either $P_{ia} < \frac{1}{2}$ (i.e. $a \succ_{\mathbf{P}} i$), or $\exists j \in [K]$ such that both $P_{ja} < \frac{1}{2}$ or $P_{ij} < \frac{1}{2}$ (i.e. $a \succ_{\mathbf{P}} j$ and $j \succ_{\mathbf{P}} i$). Thus the arm a beats every other arm under $\mathbf{P}$ either directly or via an intermediate arm, and therefore by Lemma 18, we have $a \in \mathrm{UC}(\mathbf{P})$. This establishes that SELECTPROC-UC satisfies the safe identical-arms condition w.r.t. UC.

Next, we prove that SELECTPROC-UC satisfies the safe distinct-arms property w.r.t. UC. Again, let $\mathbf{P} \in \mathcal{P}_K, \mathbf{U} \in \mathbb{R}_+^{K \times K}$ be such that $P_{ij} \leq U_{ij} \; \forall i, j$, and now suppose that SELECTPROC-UC$(\mathbf{U}) = (a, b)$ with $a \neq b$. Here it must be the case that the condition in line 7 is not satisfied. Let the set B be defined as in SELECTPROC-UC. Since $U_{ij} + U_{ji} \geq P_{ij} + P_{ji} = 1 \; \forall i, j$, by Lemma 17, we must have $B \neq \emptyset$, and therefore (a, b) must be returned via lines 11 and 15, with $b \in B$. By definition of

B, this implies that we have both $U_{ab} \geq \frac{1}{2}$ and $U_{ba} \geq \frac{1}{2}$. Thus we have

$$\begin{aligned} U_{ab} + U_{ba} &= \max(U_{ab}, U_{ba}) + \min(U_{ab}, U_{ba}) \\ &\geq \max(P_{ab}, P_{ba}) + \tfrac{1}{2} \\ &= (\tfrac{1}{2} + \Delta^{\mathbf{P}}_{ab}) + \tfrac{1}{2} \\ &= 1 + \Delta^{\mathbf{P}}_{ab}\,. \end{aligned}$$

This establishes that SELECTPROC-UC satisfies the safe distinct-arms condition w.r.t. UC (in fact, w.r.t. any tournament solution). □

D Proof of Theorem 13

PROOF. We first prove that SELECTPROC-BA satisfies the safe identical-arms condition w.r.t. BA. Let $\mathbf{P} \in \mathcal{P}_K, \mathbf{U} \in \mathbb{R}_+^{K\times K}$ be such that $P_{ij} \leq U_{ij}\ \forall i, j$, and suppose that SELECTPROC-BA$(\mathbf{U}) = (a, a)$. We will show $a \in$ BA$(\mathbf{P})$. Let the set $\mathcal{J} = \{j_1, \ldots, j_s\}$ be constructed as in SELECTPROC-BA. Now, (a, a) must be returned via line 19 of the procedure, which means the condition in line 17 must be true. In particular, this condition states that $U_{j_q,j_r} < \frac{1}{2}\ \forall 1 \leq q < r \leq s$, which implies that $P_{j_q,j_r} < \frac{1}{2}\ \forall 1 \leq q < r \leq s$. Thus the elements of $\mathcal{J}$ satisfy $j_s \succ_{\mathbf{P}} j_{s-1} \succ_{\mathbf{P}} \ldots \succ_{\mathbf{P}} j_1$. Moreover, there cannot be any arm i that beats j_s under $\mathbf{P}$, since then we would have $P_{ij} > \frac{1}{2}\ \forall j \in \mathcal{J}$ and i would have been added to $\mathcal{J}$ in line 7. Therefore, the set $\mathcal{J}$ is a true Banks trajectory under $\mathbf{P}$ (forms a maximally acyclic subtournament), and $a = j_s$ is its maximal element. Thus, $a \in$ BA$(\mathbf{P})$. This establishes that SELECTPROC-BA satisfies the safe identical-arms condition w.r.t. BA.

Next, we prove that SELECTPROC-BA satisfies the safe distinct-arms property w.r.t. BA. Again, let $\mathbf{P} \in \mathcal{P}_K, \mathbf{U} \in \mathbb{R}_+^{K\times K}$ be such that $P_{ij} \leq U_{ij}\ \forall i, j$, and now suppose that SELECTPROC-BA$(\mathbf{U}) = (a, b)$ with $a \neq b$. In this case (a, b) must be returned via line 23, and therefore, by construction, we must have both $U_{ab} \geq \frac{1}{2}$ and $U_{ba} \geq \frac{1}{2}$. Thus we have

$$\begin{aligned} U_{ab} + U_{ba} &= \max(U_{ab}, U_{ba}) + \min(U_{ab}, U_{ba}) \\ &\geq \max(P_{ab}, P_{ba}) + \tfrac{1}{2} \\ &= (\tfrac{1}{2} + \Delta^{\mathbf{P}}_{ab}) + \tfrac{1}{2} \\ &= 1 + \Delta^{\mathbf{P}}_{ab}\,. \end{aligned}$$

This establishes that SELECTPROC-BA satisfies the safe distinct-arms condition w.r.t. BA (in fact, w.r.t. any tournament solution). □

E UCB-CO: Dueling Bandit Algorithm for Copeland Set

Before describing an instantiation of our algorithmic framework designed for the Copeland set, let us briefly consider regret measures for the Copeland set:

E.1 Copeland Regret

There are many ways to define a regret measure that is normalized, symmetric, and proper w.r.t. the Copeland set; we consider one such natural measure below. In particular, for each arm $i \in [K]$, let $c_{\mathbf{P}}(i)$ denote the Copeland score of i under $\mathbf{P}$:

$$c_{\mathbf{P}}(i) = \sum_{j \neq i} \mathbf{1}(i \succ_{\mathbf{P}} j)\,,$$

Let $c^*_{\mathbf{P}}$ denote the maximal Copeland score under $\mathbf{P}$:

$$c^*_{\mathbf{P}} = \max_i c_{\mathbf{P}}(i)\,.$$

Then we define the individual Copeland regret of an arm i as its Copeland score deficit w.r.t. $c^*_{\mathbf{P}}$, normalized to lie in $[0, 1]$:

$$r^{\text{CO}}_{\mathbf{P}}(i) = \frac{c^*_{\mathbf{P}} - c_{\mathbf{P}}(i)}{c^*_{\mathbf{P}}} \quad \forall i \in [K]\,.$$

This is simply a scaled version of the Copeland regret considered by Zoghi et al. [11]. Clearly, $r^{\text{CO}}_{\mathbf{P}}(i) = 0\ \forall i \in$ CO$(\mathbf{P})$, and therefore the resulting average, weak, and strong pairwise Copeland regrets $r^{\text{CO}}_{\mathbf{P}}(\cdot, \cdot)$ are all proper w.r.t. the Copeland set (see Section 2).

Algorithm 5 SELECTPROC-CO

```
1: Input: UCB matrix U ∈ ℝ_+^{K×K}
2: for i = 1 to K do
3:     c_U(i) ← Σ_{j≠i} 1(U_ij ≥ 1/2)
4: end for
5: Select any a ∈ argmax_{i∈A} c_U(i)
6: B ← {i ≠ a : U_ai ≥ 1/2 ∧ U_ia ≥ 1/2}
7: if B = ∅ then
8:     Return (a, a)
9: else
10:    Select any b ∈ argmax_{i∈B} U_ia
11:    Return (a, b)
12: end if
```

E.2 UCB-CO Algorithm

The selection procedure SELECTPROC-CO (Algorithm 5), when instantiated in the UCB-TS template, yields the UCB-CO dueling bandit algorithm. Intuitively, SELECTPROC-CO first selects a potential Copeland winner a that beats the maximal number of other arms under $\mathbf{U}$ (lines 2–5); if there is no unresolved arm against a (line 7), then it returns (a, a) for comparison, else it selects the best-performing unresolved opponent b (line 10) and returns (a, b) for comparison. This selection procedure is quite similar to the selection procedure implicitly used in the RUCB algorithm [6]; indeed, if one were to assume the existence of a Condorcet winner, then it would be natural to look for an arm a that beats all other arms under $\mathbf{U}$ as RUCB does (rather than look for an arm a that beats a maximal number of other arms under $\mathbf{U}$ as SELECTPROC-CO does). We have the following result:

Theorem 19 (SELECTPROC-CO satisfies safety conditions w.r.t. CO). SELECTPROC-CO *satisfies both the safe identical-arms condition and the safe distinct-arms condition w.r.t.* CO.

PROOF. We first prove that SELECTPROC-CO satisfies the safe identical-arms condition w.r.t. CO. Let $\mathbf{P} \in \mathcal{P}_K, \mathbf{U} \in \mathbb{R}_+^{K\times K}$ be such that $P_{ij} \le U_{ij} \ \forall i, j$, and suppose that SELECTPROC-CO$(\mathbf{U}) = (a, a)$. We will show that $a \in \text{CO}(\mathbf{P})$. Let the set B be defined as in SELECTPROC-CO. The identical-arms pair (a, a) must be returned via line 8 of the procedure, and therefore the condition in line 7 must be satisfied, i.e. the set B must be empty. Now, consider any true Copeland winner $i^* \in \text{CO}(\mathbf{P}) \subseteq \text{TC}(\mathbf{P})$. Then $U_{i^*j} \ge P_{i^*j} > \frac{1}{2} \ \forall j \notin \text{TC}(\mathbf{P})$, and therefore

$$c_{\mathbf{U}}(i^*) \ge c_{\mathbf{P}}(i^*) \ge K - |\text{TC}(\mathbf{P})| \,.$$

Since by construction a maximizes $c_{\mathbf{U}}(\cdot)$, this implies

$$c_{\mathbf{U}}(a) \ge K - |\text{TC}(\mathbf{P})| \,.$$

Now, suppose, if possible, that $a \notin \text{TC}(\mathbf{P})$. Then there must be some $j \in \text{TC}(\mathbf{P})$ such that $U_{aj} > \frac{1}{2}$ (as otherwise, we would have $c_{\mathbf{U}}(a) \le K - |\text{TC}(\mathbf{P})| - 1$). But since $j \in \text{TC}(\mathbf{P})$, we must also then have $U_{ja} \ge P_{ja} > \frac{1}{2}$. Thus j must be unresolved against a, i.e. $j \in B$. However this contradicts the fact that B is empty; therefore, we must have $a \in \text{TC}(\mathbf{P})$. Thus we can write

$$\begin{aligned} c_{\mathbf{P}}(a) &= K - |\text{TC}(\mathbf{P})| + \sum_{j\in\text{TC}(\mathbf{P})} \mathbf{1}(P_{aj} > \tfrac{1}{2})\,; \\ c_{\mathbf{P}}(i^*) &= K - |\text{TC}(\mathbf{P})| + \sum_{j\in\text{TC}(\mathbf{P})} \mathbf{1}(P_{i^*j} > \tfrac{1}{2})\,. \end{aligned}$$

Now, we have

$$\begin{aligned} c_{\mathbf{U}}(a) &= K - |\text{TC}(\mathbf{P})| + \sum_{j\in\text{TC}(\mathbf{P})} \mathbf{1}(U_{aj} > \tfrac{1}{2}) \\ &= K - |\text{TC}(\mathbf{P})| + \sum_{j\in\text{TC}(\mathbf{P})} \mathbf{1}(P_{aj} > \tfrac{1}{2}) + \sum_{j\in\text{TC}(\mathbf{P})} \mathbf{1}(U_{aj} > \tfrac{1}{2}, P_{ja} > \tfrac{1}{2}) \\ &= c_{\mathbf{P}}(a) + \sum_{j\in\text{TC}(\mathbf{P})} \mathbf{1}(U_{aj} > \tfrac{1}{2}, P_{ja} > \tfrac{1}{2}) \\ &= c_{\mathbf{P}}(a)\,, \end{aligned}$$

since $U_{aj} > \frac{1}{2}, P_{ja} > \frac{1}{2} \implies U_{aj} > \frac{1}{2}, U_{ja} > \frac{1}{2} \implies j \in B \implies B \neq \emptyset$, which is a contradiction, and therefore we must have $\sum_{j \in \text{TC}(\mathbf{P})} \mathbf{1}(U_{aj} > \frac{1}{2}, P_{ja} > \frac{1}{2}) = 0$. Similarly,

$$\begin{aligned} c_{\mathbf{U}}(i^*) &= K - |\text{TC}(\mathbf{P})| + \sum_{j \in \text{TC}(\mathbf{P})} \mathbf{1}(U_{i^*j} > \tfrac{1}{2})\,; \\ &= K - |\text{TC}(\mathbf{P})| + \sum_{j \in \text{TC}(\mathbf{P})} \mathbf{1}(P_{i^*j} > \tfrac{1}{2}) + \sum_{j \in \text{TC}(\mathbf{P})} \mathbf{1}(U_{i^*j} > \tfrac{1}{2}, P_{ji^*} > \tfrac{1}{2}) \\ &= c_{\mathbf{P}}(i^*) + \sum_{j \in \text{TC}(\mathbf{P})} \mathbf{1}(U_{i^*j} > \tfrac{1}{2}, P_{ji^*} > \tfrac{1}{2})\,. \end{aligned}$$

This gives

$$\begin{aligned} c_{\mathbf{P}}(a) - c_{\mathbf{P}}(i^*) &= c_{\mathbf{U}}(a) - \Big(c_{\mathbf{U}}(i^*) - \sum_{j \in \text{TC}(\mathbf{P})} \mathbf{1}(U_{i^*j} > \tfrac{1}{2}, P_{ji^*} > \tfrac{1}{2}) \Big) \\ &= \big(c_{\mathbf{U}}(a) - c_{\mathbf{U}}(i^*) \big) + \sum_{j \in \text{TC}(\mathbf{P})} \mathbf{1}(U_{i^*j} > \tfrac{1}{2}, P_{ji^*} > \tfrac{1}{2}) \\ &\geq 0\,. \end{aligned}$$

Thus $c_{\mathbf{P}}(a) \geq c_{\mathbf{P}}(i^*)$, and therefore $a \in \text{CO}(\mathbf{P})$. This establishes that SELECTPROC-CO satisfies the safe identical-arms condition w.r.t. CO.

Next, we prove that SELECTPROC-CO satisfies the safe distinct-arms condition w.r.t. CO. Again, let $\mathbf{P} \in \mathcal{P}_K, \mathbf{U} \in \mathbb{R}_+^{K \times K}$ be such that $P_{ij} \leq U_{ij}\ \forall i, j$, and now suppose that SELECTPROC-CO$(\mathbf{U}) = (a, b)$ with $a \neq b$. Let the set B be defined as in SELECTPROC-CO. Now, (a, b) must be returned via line 11, and therefore we must have $b \in B$. By definition of B, this implies that we have both $U_{ab} \geq \frac{1}{2}$ and $U_{ba} \geq \frac{1}{2}$. Thus we have

$$\begin{aligned} U_{ab} + U_{ba} &= \max(U_{ab}, U_{ba}) + \min(U_{ab}, U_{ba}) \\ &\geq \max(P_{ab}, P_{ba}) + \tfrac{1}{2} \\ &= (\tfrac{1}{2} + \Delta_{ab}^{\mathbf{P}}) + \tfrac{1}{2} \\ &= 1 + \Delta_{ab}^{\mathbf{P}}\,. \end{aligned}$$

This establishes that SELECTPROC-CO satisfies the safe distinct-arms condition w.r.t. CO (in fact, w.r.t. any tournament solution). □

By virtue of Theorem 8, this immediately yields the following regret bound for UCB-CO (as for our other algorithms, the regret bound holds for all regret measures that are normalized, symmetric, and proper w.r.t. the Copeland set; here we apply it to the pairwise Copeland regret measures resulting from the individual Copeland regret defined above):

Corollary 20 (**Regret bound for UCB-CO algorithm**). *Let $\mathbf{P} \in \mathcal{P}_K$. Let $\alpha > \frac{1}{2}$ and $\delta \in (0, 1]$. Then with probability at least $1 - \delta$, the cumulative regret of UCB-CO w.r.t. the Copeland set satisfies*

$$\mathcal{R}_T^{\text{CO}}\big(\text{UCB-CO}(\alpha)\big) \;\leq\; C(K, \alpha, \delta) + 4\alpha\,(\ln T) \Bigg(\sum_{i<j : (i,j) \notin \text{CO}(\mathbf{P}) \times \text{CO}(\mathbf{P})} \frac{r_{\mathbf{P}}^{\text{CO}}(i, j)}{(\Delta_{ij}^{\mathbf{P}})^2} \Bigg)\,.$$

F Supplement to Section 5 (Experiments)

Below we provide details of the preference matrices used in our experiments (Section F.1), and give complete experimental results (Section F.2).

F.1 Preference Matrices Used in Our Experiments

We used three preference matrices in our experiments: $\mathbf{P}^{\text{Hudry}}$, $\mathbf{P}^{\text{Tennis}}$, and $\mathbf{P}^{\text{MSLR}}$. These matrices are described below.

F.1.1 $\mathbf{P}^{\text{Hudry}}$

The Hudry tournament, shown in Figure 2(b), is a well-studied tournament on 13 nodes, and has the special property that it is the smallest tournament for which the Banks and Copeland sets are disjoint [14]. As seen in Figure 2(b), the Hudry tournament has a Copeland set of size 1, a Banks set of size 3, an uncovered set of size 4 (containing both the Copeland set and the Banks set), and a top cycle of size 13 (i.e. containing all 13 nodes). We constructed the $\mathbf{P}^{\text{Hudry}}$ preference matrix so that its induced tournament corresponded to the Hudry tournament; the pairwise probabilities were designed to give the Copeland winner only a small margin over its dominion (arms it beats), while other arms, in particular the Banks winners, had higher margins over their dominion:

$$\mathbf{P}^{\text{Hudry}} = \begin{bmatrix} 0.5 & 0.1 & 0.1 & 0.1 & 0.6 & 0.6 & 0.6 & 0.6 & 0.6 & 0.6 & 0.6 & 0.6 & 0.6 \\ 0.9 & 0.5 & 0.9 & 0.1 & 0.1 & 0.1 & 0.1 & 0.9 & 0.9 & 0.9 & 0.9 & 0.9 & 0.9 \\ 0.9 & 0.1 & 0.5 & 0.9 & 0.9 & 0.9 & 0.9 & 0.1 & 0.1 & 0.1 & 0.9 & 0.9 & 0.9 \\ 0.9 & 0.9 & 0.1 & 0.5 & 0.9 & 0.9 & 0.9 & 0.9 & 0.9 & 0.9 & 0.1 & 0.1 & 0.1 \\ 0.4 & 0.9 & 0.1 & 0.1 & 0.5 & 0.9 & 0.9 & 0.9 & 0.9 & 0.9 & 0.1 & 0.1 & 0.1 \\ 0.4 & 0.9 & 0.1 & 0.1 & 0.1 & 0.5 & 0.9 & 0.9 & 0.9 & 0.9 & 0.1 & 0.1 & 0.1 \\ 0.4 & 0.9 & 0.1 & 0.1 & 0.1 & 0.1 & 0.5 & 0.9 & 0.9 & 0.9 & 0.1 & 0.1 & 0.1 \\ 0.4 & 0.1 & 0.9 & 0.1 & 0.1 & 0.1 & 0.1 & 0.5 & 0.9 & 0.9 & 0.9 & 0.9 & 0.9 \\ 0.4 & 0.1 & 0.9 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.5 & 0.9 & 0.9 & 0.9 & 0.9 \\ 0.4 & 0.1 & 0.9 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.5 & 0.9 & 0.9 & 0.9 \\ 0.4 & 0.1 & 0.1 & 0.9 & 0.9 & 0.9 & 0.9 & 0.1 & 0.1 & 0.1 & 0.5 & 0.9 & 0.9 \\ 0.4 & 0.1 & 0.1 & 0.9 & 0.9 & 0.9 & 0.9 & 0.1 & 0.1 & 0.1 & 0.1 & 0.5 & 0.9 \\ 0.4 & 0.1 & 0.1 & 0.9 & 0.9 & 0.9 & 0.9 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.5 \end{bmatrix}$$

F.1.2 $\mathbf{P}^{\text{Tennis}}$

We constructed the $\mathbf{P}^{\text{Tennis}}$ preference matrix by compiling the all-time win-loss results of tennis matches among 8 international tennis players as recorded by the Association of Tennis Professionals (ATP).[8] In particular, we considered matches among the following 8 players:

1	Goran Ivanisevic
2	Stefan Edberg
3	Pete Sampras
4	Boris Becker
5	Andre Agassi
6	Ivan Lendl
7	Michael Chang
8	Jim Courier

For each pair of players i and j, we took P_{ij}^{Tennis} to be the fraction of matches between i and j that were won by i. This resulted in the following pairwise preference matrix:

$$\mathbf{P}^{\text{Tennis}} = \begin{bmatrix} 0.50 & 0.47 & 0.67 & 0.53 & 0.57 & 0.83 & 0.55 & 0.73 \\ 0.53 & 0.50 & 0.57 & 0.71 & 0.67 & 0.48 & 0.43 & 0.60 \\ 0.33 & 0.43 & 0.50 & 0.37 & 0.41 & 0.38 & 0.40 & 0.20 \\ 0.47 & 0.29 & 0.63 & 0.50 & 0.71 & 0.52 & 0.17 & 0.14 \\ 0.43 & 0.33 & 0.59 & 0.29 & 0.50 & 0.75 & 0.32 & 0.58 \\ 0.17 & 0.52 & 0.62 & 0.48 & 0.25 & 0.50 & 0.29 & 0.00 \\ 0.45 & 0.57 & 0.60 & 0.83 & 0.68 & 0.71 & 0.50 & 0.52 \\ 0.27 & 0.40 & 0.80 & 0.86 & 0.42 & 1.00 & 0.48 & 0.50 \end{bmatrix}$$

The tournament associated with $\mathbf{P}^{\text{Tennis}}$ is shown in Figure 2(c). As can be seen, this tournament has a large top cycle of 7 players (all players except Pete Sampras). The uncovered set and Banks set here are identical, and contain 3 players (Goran Ivanisevic, Stefan Edberg, and Michael Chang); of these, only 2 players (Goran Ivanisevic and Michael Chang) are in the Copeland set.

F.1.3 $\mathbf{P}^{\text{MSLR}}$

The Microsoft Learning to Rank (MSLR) Web10K data set contains 10,000 web-search query and document pairs, each associated with 132 features; each query-document pair is labeled with a user-assigned relevance score.[9] Following the procedure adopted by Jamieson and Nowak [16], one can treat the 132 query-document features as arms, and can consider pairwise comparisons among the features/arms in terms of how well they rank pairs of documents for a given query (judged by the user-provided relevance scores). Specifically, in order to draw a pairwise comparison between features f_i and f_j, one would randomly sample a query q and two associated documents d and d', and and would test whether one feature is better than the other in terms of ranking d and d' relative to the user-assigned scores $s(q,d), s(q,d')$: if $(f_i(q,d) - f_i(q,d'))(s(q,d) - s(q,d')) > 0$ and $(f_j(q,d) - f_j(q,d'))(s(q,d) - s(q,d')) < 0$, then feature f_i wins over feature f_j; if $(f_i(q,d) -$

[8] ATP website: `http://www.atpworldtour.com`

[9] This data set is available from: `http://research.microsoft.com/en-us/projects/mslr/`

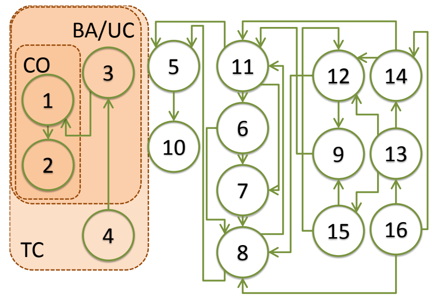


Figure 5: Tournament associated with the $\mathbf{P}^{\text{MSLR}}$ preference matrix together with its tournament solutions. As in Figure 2, edges that are not explicitly shown are directed from left to right.

$f_i(q, d'))(s(q, d) - s(q, d')) < 0$ and $(f_j(q, d) - f_j(q, d'))(s(q, d) - s(q, d')) > 0$, then feature f_j wins over feature f_i; and otherwise, there is a tie. Thus, for any pair of features f_i and f_j, one can estimate the associated pairwise preference probability P_{ij} by sampling a few queries q and associated document pairs d, d' and counting the fraction of times f_i wins over f_j (adjusting for ties by counting half a win for each tie).

In our experiments, we used a subset of 16 features, and constructed a preference matrix $\mathbf{P}^{\text{MSLR}} \in \mathcal{P}_{16}$ by randomly sampling, for each of the $\binom{16}{2}$ pairs of features, 25 queries and document pairs as above, and counting the fraction of wins for each pair. This resulted in the following preference matrix:[10]

$$\mathbf{P}^{\text{MSLR}} = \begin{bmatrix}
0.50 & 0.58 & 0.48 & 0.52 & 0.56 & 0.64 & 0.54 & 0.56 & 0.62 & 0.54 & 0.54 & 0.66 & 0.52 & 0.58 & 0.62 & 0.56 \\
0.42 & 0.50 & 0.72 & 0.60 & 0.56 & 0.66 & 0.56 & 0.64 & 0.56 & 0.64 & 0.52 & 0.68 & 0.56 & 0.54 & 0.60 & 0.54 \\
0.52 & 0.28 & 0.50 & 0.48 & 0.68 & 0.54 & 0.52 & 0.52 & 0.72 & 0.68 & 0.58 & 0.60 & 0.52 & 0.64 & 0.64 & 0.72 \\
0.48 & 0.40 & 0.52 & 0.50 & 0.54 & 0.54 & 0.54 & 0.62 & 0.66 & 0.54 & 0.58 & 0.58 & 0.52 & 0.62 & 0.64 & 0.62 \\
0.44 & 0.44 & 0.32 & 0.46 & 0.50 & 0.58 & 0.60 & 0.48 & 0.54 & 0.54 & 0.48 & 0.60 & 0.68 & 0.56 & 0.64 & 0.52 \\
0.36 & 0.34 & 0.46 & 0.46 & 0.42 & 0.50 & 0.64 & 0.60 & 0.66 & 0.46 & 0.48 & 0.54 & 0.58 & 0.52 & 0.72 & 0.64 \\
0.46 & 0.44 & 0.48 & 0.46 & 0.40 & 0.36 & 0.50 & 0.52 & 0.66 & 0.48 & 0.42 & 0.54 & 0.62 & 0.58 & 0.52 & 0.52 \\
0.44 & 0.36 & 0.48 & 0.38 & 0.52 & 0.40 & 0.48 & 0.50 & 0.54 & 0.44 & 0.58 & 0.48 & 0.58 & 0.52 & 0.60 & 0.48 \\
0.38 & 0.44 & 0.28 & 0.34 & 0.46 & 0.34 & 0.34 & 0.46 & 0.50 & 0.42 & 0.52 & 0.48 & 0.58 & 0.52 & 0.48 & 0.52 \\
0.46 & 0.36 & 0.32 & 0.46 & 0.46 & 0.54 & 0.52 & 0.56 & 0.58 & 0.50 & 0.56 & 0.58 & 0.74 & 0.60 & 0.64 & 0.52 \\
0.46 & 0.48 & 0.42 & 0.42 & 0.52 & 0.52 & 0.58 & 0.42 & 0.48 & 0.44 & 0.50 & 0.60 & 0.64 & 0.48 & 0.56 & 0.54 \\
0.34 & 0.32 & 0.40 & 0.42 & 0.40 & 0.46 & 0.46 & 0.52 & 0.52 & 0.42 & 0.40 & 0.50 & 0.48 & 0.48 & 0.48 & 0.52 \\
0.48 & 0.44 & 0.48 & 0.48 & 0.32 & 0.42 & 0.38 & 0.42 & 0.42 & 0.26 & 0.36 & 0.52 & 0.50 & 0.54 & 0.56 & 0.48 \\
0.42 & 0.46 & 0.36 & 0.38 & 0.44 & 0.48 & 0.42 & 0.48 & 0.48 & 0.40 & 0.52 & 0.52 & 0.46 & 0.50 & 0.46 & 0.42 \\
0.38 & 0.40 & 0.36 & 0.36 & 0.36 & 0.28 & 0.48 & 0.40 & 0.52 & 0.36 & 0.44 & 0.52 & 0.44 & 0.54 & 0.50 & 0.52 \\
0.44 & 0.46 & 0.28 & 0.38 & 0.48 & 0.36 & 0.48 & 0.52 & 0.48 & 0.48 & 0.46 & 0.48 & 0.52 & 0.58 & 0.48 & 0.50
\end{bmatrix}$$

The tournament associated with $\mathbf{P}^{\text{MSLR}}$ is shown in Figure 5. As can be seen, this tournament has a small top cycle of size 4, identical Banks and uncovered sets of size 3, and a Copeland set of size 2.

F.2 Complete Experimental Results

Complete results of our experiments on the three preference matrices above are shown in Figures 6, 7, and 8, respectively. As noted in Section 5, all algorithms were assessed on average pairwise regret relative to the target tournament solution of interest, as defined in Section 2 and Appendix E. The plots show regret performance averaged over 10 independent runs, with light colored bands representing one standard error.

As can be seen, in most cases, our dueling bandit algorithms outperform existing baselines in terms of minimizing regret relative to the tournament solutions of interest. For the Copeland regret, our UCB-CO algorithm performs similarly in practice to the CCB algorithm. The SAVAGE-CO algorithm, due to its use of the confidence parameter $\delta = 1/T$ to ensure a meaningful regret bound, tends to require a large number of trials for exploration (seen in our plots as an initial high-regret period), before turning sharply to exploitation (seen as an abrupt change to a near-zero additional regret phase).

[10] Occasionally, preference matrices generated by sampling from real-world data may contain "tied" preferences, i.e. pairs $i \neq j$ with empirically observed $P_{ij} = P_{ji} = \frac{1}{2}$. In such cases, we suggest breaking ties by adding small random perturbations to P_{ij} and P_{ji} using the minimal margin between non-tied arms, defining $P'_{ij} = \frac{1}{2} + \epsilon\Delta_{\text{min}}$ and $P'_{ji} = \frac{1}{2} - \epsilon\Delta_{\text{min}}$, where $\Delta_{\min} = \min_{(i,j):P_{ij} \neq \frac{1}{2}} \Delta_{ij}$ and ϵ is a Rademacher random variable taking values in $\{\pm 1\}$ with equal probability. This results in a tie-free perturbed matrix $\mathbf{P}'$ which retains the same minimal margin as the original tied matrix $\mathbf{P}$.

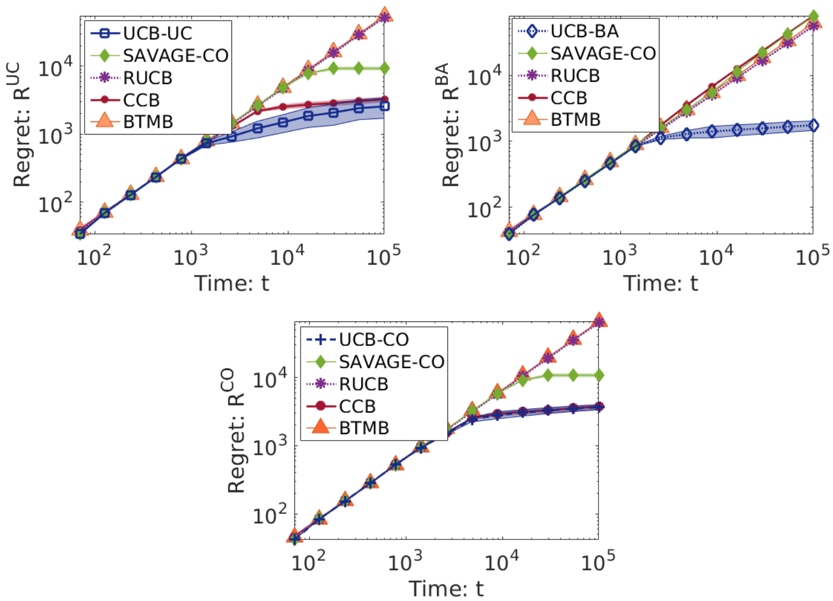


Figure 6: Regret performance on (pairwise comparison outcomes from) $\mathbf{P}^{\text{Hudry}}$. **Top left:** Uncovered set regret. **Top right:** Banks set regret. **Bottom:** Copeland set regret. (Note: We do not consider top cycle regret for $\mathbf{P}^{\text{Hudry}}$ since in this case the top cycle is the entire set of arms.)

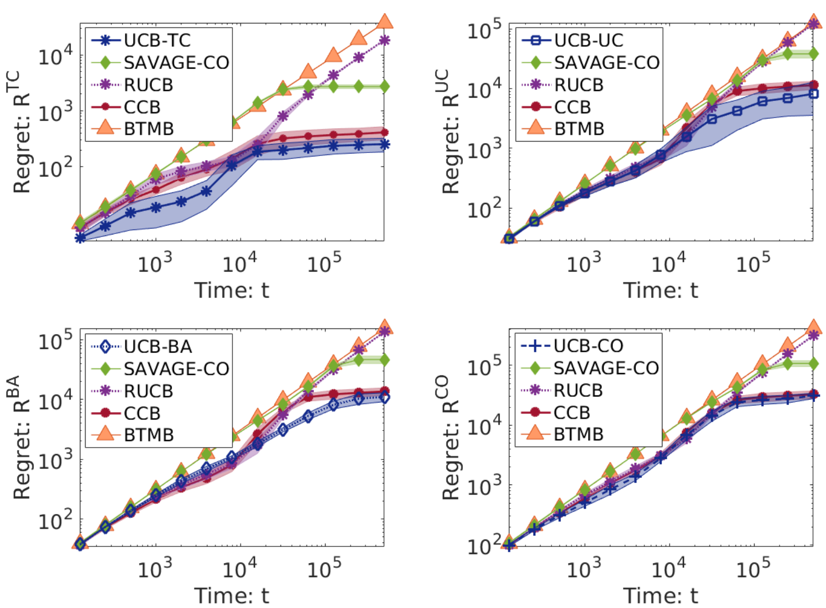


Figure 7: Regret performance on (pairwise comparison outcomes from) $\mathbf{P}^{\text{Tennis}}$. **Top left:** Top cycle regret. **Top right:** Uncovered set regret. **Bottom left:** Banks set regret. **Bottom right:** Copeland set regret.

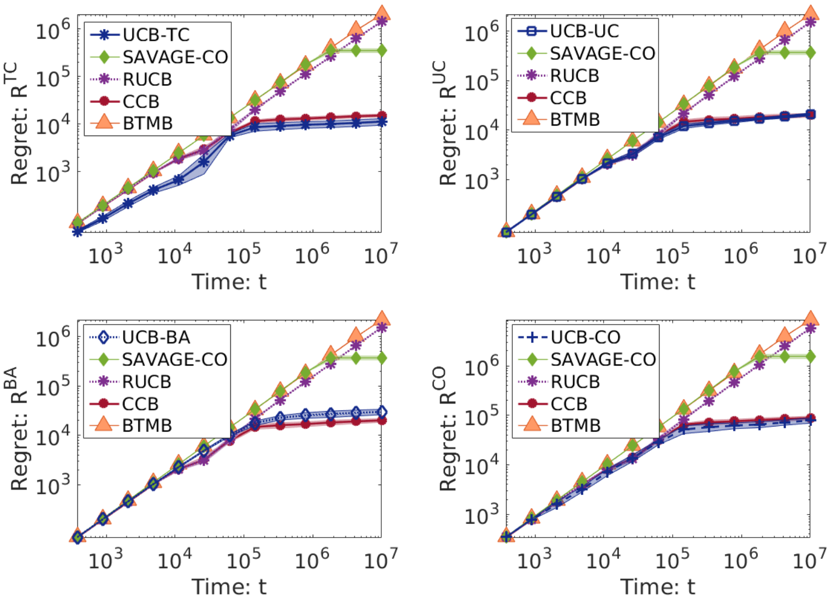


Figure 8: Regret performance on (pairwise comparison outcomes from) $\mathbf{P}^{\text{MSLR}}$. **Top left:** Top cycle regret. **Top right:** Uncovered set regret. **Bottom left:** Banks set regret. **Bottom right:** Copeland set regret.