Copeland Dueling Bandit Problem: Regret Lower Bound, Optimal Algorithm, and Computationally Efficient Algorithm

Junpei Komiyama JUNPEI@KOMIYAMA.INFO
The University of Tokyo, Japan

Junya Honda HONDA@STAT.T.U-TOKYO.AC.JP
The University of Tokyo, Japan

Hiroshi Nakagawa NAKAGAWA@DL.ITC.U-TOKYO.AC.JP
The University of Tokyo, Japan

Abstract

We study the K-armed dueling bandit problem, a variation of the standard stochastic bandit problem where the feedback is limited to relative comparisons of a pair of arms. The hardness of recommending Copeland winners, the arms that beat the greatest number of other arms, is characterized by deriving an asymptotic regret bound. We propose Copeland Winners Relative Minimum Empirical Divergence (CW-RMED) and derive an asymptotically optimal regret bound for it. However, it is not known whether the algorithm can be efficiently computed or not. To address this issue, we devise an efficient version (ECW-RMED) and derive its asymptotic regret bound. Experimental comparisons of dueling bandit algorithms show that ECW-RMED significantly outperforms existing ones.

1. Introduction

A multi-armed bandit problem is a crystallized instance of a sequential decision-making problem in an uncertain environment, and it can model many real-world scenarios. This problem involves conceptual entities called arms. At each round, the forecaster draws one of the K arms and receives a corresponding reward feedback. The aim of the forecaster is to maximize the cumulative reward over rounds, which is achieved by running an algorithm that balances the exploration (acquisition of information) and the exploitation (utilization of information). In evaluating the performance of a bandit algorithm, a metric called regret, which measures how much the algorithm explores, is widely used.

While it is desirable to obtain rewards as direct feedback from an arm, in a number of practical cases such direct feedback is not available. In this paper, we consider a version of the standard stochastic bandit problem called the K-armed dueling bandit problem (Yue et al., 2009), in which the forecaster receives relative feedback, which specifies which of the two arms is preferred. Although the original motivation of the dueling bandit problem arose in the field of information retrieval, learning under relative feedback is universal to many fields, such as recommender systems (Gemmis et al., 2009), graphical design (Brochu et al., 2010), and natural language processing (Zaidan & Callison-Burch, 2011), which involve explicit or implicit feedback provided by humans.

In the standard bandit problem, the best arm is naturally defined as the one with the largest expected reward. However, if the feedback is restricted to the results of pairwise comparisons, there are several possible ways to define the best arm. Following the literature on the dueling bandit problem, we call the best arm the winner. When there exists an arm that beats (i.e., preferred in expectation) all the other arms, it is natural to define it as the winner; this notion is called a Condorcet winner. Unfortunately, the Condorcet winner does not always exist. Still, we can define an extended notion of the Condorcet winner that always exists as follows. Let the Copeland winners be the arms that beat the greatest number of other arms. In this paper, we study the difficulty of finding the Copeland winners from pairwise feedback.

1.1. Related work

Early algorithms for solving the dueling bandit problem, such as Interleaved Filter (Yue et al., 2012) and Beat the Mean Bandit (Yue & Joachims, 2011), require the arms to

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.

be totally ordered.

Urvoy et al. (2013) considered a large class of sequential learning problems that includes the dueling bandit problem and introduced the notion of Condorcet, Copeland, and Borda dueling bandit problems. Several algorithms, such as Relative Upper Confidence Bound (RUCB) (Zoghi et al., 2014), and Relative Minimum Empirical Divergence (RMED) (Komiyama et al., 2015a), have since been proposed that effectively solve the Condorcet dueling bandit problem. The assumption on the Condorcet winner partly relaxes the assumption of the total order because it admits circular preferences that involve non-winners.

However, as the Condorcet winner does not always exist, the result of running one of these algorithms is unpredictable if it is applied to an instance without a Condorcet winner. Consequently, the practical applicability of Condorcet dueling bandit algorithms is limited. Some papers have discussed the problem of preference elicitation without a Condorcet winner (Jamieson et al., 2015; Zoghi et al., 2015a) and have motivated studies of more general dueling bandit problems. Unlike the Condorcet winner, the Borda and Copeland winners always exist. Note that there are also other notions of winners, such as the von Neumann winner (Dudík et al., 2015) or Random walk winner (Altman & Tennenholtz, 2008) together with their corresponding dueling bandit problems. Among them, we will consider the Copeland dueling bandit problem. Unlike the Borda or Random Walk winners, the Copeland winners are compatible with the Condorcet winner; if the Condorcet winner exists, it is also the Copeland winner. An algorithm for finding the Copeland winner (i) covers the application range of the Condorcet winner and (ii) can find arms that beat other arms the most, even if the Condorcet winner does not exist.

Another line of study is on the partial monitoring problem (Bartók et al., 2014). The partial monitoring is general enough to cover the multi-armed bandit. Some classes of dueling bandit problems, such as utility-based ones (Gajane & Urvoy, 2015), can also be formalized as a partial monitoring. However, it is unknown as to whether the Copeland dueling bandit problem can be effectively represented as a partial monitoring or not. Moreover, existing algorithms for partial monitoring, such as Bayes-update Partial Monitoring (BPM) (Vanchinathan et al., 2014) or Partial Monitoring Deterministic Minimum Empirical Divergence (PM-DMED) (Komiyama et al., 2015b), are not very scalable to the number of actions.

Existing results on the Copeland dueling bandit problem: The difficulty of the dueling bandit problem lies in that there are $O(K^2)$ pairs. There are some algorithms, such as Sensitivity Analysis of VAriables for Generic Exploration (SAVAGE) (Urvoy et al., 2013), Preference-based Racing (PBR) (Busa-Fekete et al., 2013), and Rank Elicitation (RankEI) (Busa-Fekete et al., 2014), that can deal with general classes of problems that entail solving Copeland dueling bandit problems. The price to pay for such generality is performance: all three algorithms have $O(K^2 \log T)$ regret because they naively compare all pairs $O(\log T)$ times.

The recently proposed Copeland Confidence Bound (CCB) (Zoghi et al., 2015a) exploits the structure of the Copeland dueling bandit problem and is relatively efficient. It has an asymptotic regret of $O(\frac{K(C+L_1+1)}{\Delta^2} \log T)$ (Theorem 3 in Zoghi et al. 2015a), where C is the number of Copeland winners, L_1 is the number of arms that beats the Copeland winner, and Δ is related to how hard it is to determine whether each arm i beats arm j or not. In this paper, we further push our understanding of the dueling bandit problem by deriving an asymptotically optimal regret bound. The optimal bound states that (i) the dependency on C can be completely removed; (ii) the dependency on L_1 is necessary for some cases but unnecessary for typical cases explained later, and (iii) the dependency on Δ can be relaxed by introducing a divergence-based bound. In an information retrieval example, the optimal bound improves the one of the CCB by several orders of magnitude (Table 1).

Contributions: The main contributions of this paper are summarized in the following four aspects: First, we derive an asymptotic regret lower bound (Section 3). The lower bound is based on the minimum amount of exploration for identifying a Copeland winner. Second, we propose the Copeland Winners Relative Minimum Empirical Divergence (CW-RMED) algorithm. CW-RMED is the first algorithm whose performance asymptotically matches the regret lower bound (Section 4.1). Unfortunately, a naive implementation of CW-RMED is computationally prohibitive. Third, we propose Efficient Copeland Winners RMED (ECW-RMED), another algorithm that addresses the above computational issue (Section 4.2). An efficient way to implement it is proposed. Moreover, we show that the regret of ECW-RMED is very close to optimal. Finally, we implemented ECW-RMED and compared its performance with those of existing algorithms (Section 5). ECW-

Table 1. Comparison of leading logarithmic constants of regret bounds on the Microsoft Learning to Rank dataset. CW-RMED and ECW-RMED are the algorithms proposed in this paper. The values are averaged over 10^4 randomly generated submatrices of size 16×16. Details of the dataset are presented in Section 5. The bound of CCB is about 1000 times looser than the optimal.

Optimal: CW-RMED	ECW-RMED	CCB (Zoghi et al., 2015a)
6.7×10^2	7.3×10^2	8.8×10^5

RMED significantly outperformed the state-of-the-art algorithms on many datasets. In a ranker evaluation example, its regret was smaller than one third of those of the others.

2. Problem Setup

The K-armed dueling bandit problem involves K arms that are indexed as $[K] := \{1, 2, \dots, K\}$. Let $M \in \mathbb{R}^{K \times K}$ be a preference matrix whose ij entry $\mu_{i,j}$ corresponds to the probability that arm i is preferred to arm j. At each round $t = 1, 2, \dots, T$, the forecaster draws a pair of arms $p(t) = (l(t), m(t)) \in [K]^2$ and, receives relative feedback $\hat{X}_{l(t),m(t)}(t) \sim \mathrm{Bernoulli}(\mu_{l(t),m(t)})$ that indicates which of $(l(t), m(t))$ is preferred. We say arm i beats arm j if $\mu_{i,j} > 1/2$. By definition, $\mu_{i,j} = 1 - \mu_{j,i}$ holds for any $i, j \in [K]$ and $\mu_{i,i} = 1/2$. Throughout this paper, we assume $\mu_{i,j} \neq 1/2$ for $i \neq j$. Let $\mathcal{P}_{i \neq j} := \{(i, j) : i, j \in [K], i > j\}$ and $\mathcal{P}_{\mathrm{all}} := \{(i, j) : i, j \in [K], i \geq j\}$. A comparison of pair (i, j) is identified with that of pair (j, i).

Let $N_{i,j}(t)$ be the number of comparisons of pair (i, j) and $\hat{\mu}_{i,j}(t)$ be the empirical estimate of $\mu_{i,j}$ at round t. For $j \neq i$, let $N_{i>j}(t)$ be the number of times i is preferred over j. Accordingly, $\hat{\mu}_{i,j}(t) = N_{i>j}(t)/N_{i,j}(t)$, where we set $0/0 = 1/2$ here.

Let the superiors of arm i be $S_i := \{j : j \in [K], \mu_{i,j} < 1/2\}$, that is, the set of arms that beat arm i. Let $L_i := |S_i|$ and $C = |\{i \in [K] : L_i = \min_j L_j\}|$. Without loss of generality, we can assume $L_1 = L_2 = \cdots = L_C \leq \cdots \leq L_K$. Of course, algorithms should not exploit this ordering. Arms $[C]$ are called Copeland winners. Note that the Copeland winners always exist, but are not necessarily unique. Let the inferiors of arm i be $I_i := \{j : j \in [K], \mu_{i,j} > 1/2\}$. Assuming that $\mu_{i,j} \neq 1/2$ for $i \neq j$, each arm j is either a superior or an inferior of arm i. When $L_1 = 0$, the Copeland winner is unique and also called a Condorcet winner.

We define the regret per round[1] is $r_{i,j} := (L_i + L_j - 2L_1)/(2(K-1)) \leq 1$ when the pair (i, j) is compared and the regret as $R(T) := \sum_{t \in [T]} r_{l(t),m(t)}$. The regret increases at each round unless both $l(t)$ and $m(t)$ are Copeland winners. This definition is reasonable because we have defined the goodness of an arm by the number of arms that i beats (Copeland number) and are interested in drawing the best arms. The choice of $l(t) = m(t)$ is possible, but yields no useful information since, by definition, $\mu_{i,i} = 1/2$ for any arm i.

Note that, we can also consider other definitions of regret; the analysis in this paper is relied on the facts that regret per round $r_{i,j}$ is (i) finite, (ii) determined by the Copeland numbers, (iii) and equal to zero if i and j are Copeland winners. For example, we can consider a regret such that $r_{i,j} = 0$ if $i, j \in [C]$ and 1 otherwise, and easily modify our result in accordance with that definition.

[1]The constant factor of this definition is different from the one defined in Zoghi et al. (2015a). Our result can be compared with that of Zoghi et al. (2015a) simply by multiplying a constant.

3. Regret Lower Bound

In this section, we derive an asymptotic regret lower bound when $T \to \infty$. In the context of the standard multi-armed bandit problem, Lai & Robbins (1985) derived the regret lower bound of strongly consistent algorithms; intuitively, a strongly consistent algorithm is "uniformly good" in the sense that it works well with any set of model parameters. We extend this result to the Copeland dueling bandit problem.

We first define notions that are important in characterizing the regret lower bound: the subsets of the power set of the superiors and the inferiors with a fixed size. Let $\mathcal{S}_i^m := \{S \in 2^{S_i} : |S| = m\}$, $\mathcal{I}_i^m := \{I \in 2^{I_i} : |I| = m\}$, and $\mathcal{S}_i^{\setminus j,m} := \{S \in 2^{S_i \setminus \{j\}} : |S| = m\}$. Moreover, let $\mathcal{M}_{\mathrm{Cop}}$ be a set of all preference matrices of size $K \times K$. A Copeland dueling bandit algorithm is strongly consistent if it satisfies $\mathbb{E}[R(T)] = o(T^a)$ for any $a > 0$ given any preference matrix $M \in \mathcal{M}_{\mathrm{Cop}}$. Essentially, a strongly consistent algorithm needs to find one of the Copeland winners with a high confidence level. To make sure that arm i^* is a Copeland winner, we need to simultaneously find (i) an upper-bound L_{i^*} of a Copeland winner i^* and (ii) a lower-bound L_j of the other arms. The minimum amount of exploration in Copeland dueling bandit is characterized in this way. The following lemma formalizes the aforementioned statement.

Lemma 1. *(Lower bound on the number of draws) Let $d_{\mathrm{KL}}(p, q) := p \log p/q + (1-p) \log (1-p)/(1-q)$ be the Kullback-Leibler (KL) divergence between two Bernoulli distributions with parameters p, q. For any strongly consistent algorithm, the following inequality holds for at least one $i_1 \in [C]$:*

$$\begin{aligned} &\forall_{i_2 \neq i_1} \forall l \in \{\max\{0, L_1 - 1\}, \dots, L_2\} \\ &\quad \forall I \in \mathcal{I}_{i_1}^{l+1-L_{i_1}} \forall S \in \mathcal{S}_{i_2}^{\setminus i_1, \max\{0, L_{i_2} - l - \mathbf{1}\{i_2 \in I\}\}} \\ &\sum_{(i,j) \in \mathcal{P}_{IS}} d_{\mathrm{KL}}(\mu_{i,j}, 1/2) \mathbb{E}\left[N_{i,j}(T)\right] \geq (1 - o(1)) \log T, \end{aligned} \tag{1}$$

where

$$\begin{aligned} \mathcal{P}_{IS} &= \mathcal{P}_{IS}(i_1, i_2, l, I, S) \\ &:= \{(i_1, j) : j \in I\} \cup \{(i_2, j) : j \in S\}. \end{aligned}$$

Intuitively, Lemma 1 can be interpreted as follows: for each round t, consistency requires an algorithm to identify one of the Copeland winners i_1 with confidence level

$1/t$. For some i_2, l, I, and S, if the preferences among the pairs in $\mathcal{P}_{IS}(i_1, i_2, l, I, S)$ are inverted, then arm $i_2 \neq i_1$ has $L_{i_2} \leq l$ and $L_{i_1} \geq l+1$, which implies that i_1 is not a Copeland winner. We need to limit all such risks for all possible i_2, l, I, S. Each risk is calculated in accordance with the large deviation principle (Cover & Thomas, 2006) as $\sim \exp\left(-\sum_{\mathcal{P}_{IS}} d_{\mathrm{KL}}(\mu_{i,j}, 1/2) N_{i,j}(t)\right)$, and the algorithm must continue comparing pairs in $\mathcal{P}_{IS}$ until $\sum_{\mathcal{P}_{IS}} d_{\mathrm{KL}}(\mu_{i,j}, 1/2) N_{i,j}(t) \sim \log t$ in order to lower the risk to $\exp(-\log t) = 1/t$ for each $\mathcal{P}_{IS}$.

The proof of Lemma 1 is in Appendix E. Note that all proofs are in Supplementary Material. The technique used in the proof extends the one of Lai & Robbins (1985) for the standard bandit problem in two aspects: (i) in the standard bandit problem, each arm is associated with a single distribution, whereas in the Copeland dueling bandit problem each arm is related to $K-1$ distributions (i.e., comparison with other arms). Therefore, not all of the pairs are required to be drawn, and we need a sophisticated analysis to determine the set of conditions that consistency requires. Moreover, (ii) the Copeland winner is not necessarily unique; there can be several ties with the maximum Copeland number. We show that consistency requires an algorithm to find at least one of the Copeland winners, but it does not need to find all of them.

Next, we derive the asymptotic regret lower bound, which is the minimum amount of regret such that inequality (1) is satisfied. Let $\{\nu_{i,j}\}$ be a $K \times K$ preference matrix, and let the superiors and the inferiors under the preference matrix $\{\nu_{i,j}\}$ be $\hat{S}_i = \hat{S}_i(\{\nu_{i,j}\}) := \{j \in [K] : \nu_{i,j} < 1/2\}$ and $\hat{I}_i = \hat{I}_i(\{\nu_{i,j}\}) := \{j \in [K] : \nu_{i,j} > 1/2\}$. Moreover, let the number of the superiors be $\hat{L}_i = \hat{L}_i(\nu_{i,j}) := |\hat{S}_i(\nu_{i,j})|$, and the a-th smallest element among $\{\hat{L}_i\}_{i\in[K]}$ be $\hat{L}^{(a)} = \hat{L}^{(a)}(\{\nu_{i,j}\})$; let the Copeland winner be $\hat{\mathcal{C}}_{\mathrm{cop}} = \hat{\mathcal{C}}_{\mathrm{cop}}(\{\nu_{i,j}\}) := \{i : \hat{L}_i = \hat{L}^{(1)}(\{\nu_{i,j}\})\} \subset [K]$. $\hat{\mathcal{S}}_i^m(\{\nu_{i,j}\})$, $\hat{\mathcal{I}}_i^m(\{\nu_{i,j}\})$, and $\hat{\mathcal{S}}_i^{\setminus j,m}(\{\nu_{i,j}\})$ are defined in the same way. For $i_1 \in \hat{\mathcal{C}}_{\mathrm{cop}}$, let

$$\mathcal{R}_{i_1}(\{\nu_{i,j}\}) := \Bigg\{ \{q_{i,j}\}_{i>j} \in [0, 1/d_{\mathrm{KL}}(\nu_{i,j}, 1/2)]^{K(K-1)/2} : \forall i_2 \neq i_1 \; \forall l \in \{\max\{0, \hat{L}^{(1)} - 1\}, \ldots, \hat{L}^{(2)}\} \; \forall I \in \hat{\mathcal{I}}_{i_1}^{(l+1-\hat{L}^{(1)})} \; \forall S \in \hat{\mathcal{S}}_{i_2}^{\setminus i_1, \max\{0, \hat{L}_{i_2} - l - \mathbf{1}\{i_2 \in I\}\}} \sum_{(i,j)\in\mathcal{P}_{IS}} q_{i,j} d_{\mathrm{KL}}(\nu_{i,j}, 1/2) \geq 1 \Bigg\}. \quad (2)$$

Note that $\mathcal{R}_{i_1}(\{\nu_{i,j}\})$ is non-empty because it includes a trivial solution $q_{i,j} = 1/d_{\mathrm{KL}}(\nu_{i,j}, 1/2)$ for each $(i,j) \in \mathcal{P}_{i\neq j}$. Moreover, let $\hat{r}_{i,j}(\{\nu_{i,j}\}) := (\hat{L}_i + \hat{L}_j - 2\hat{L}^{(1)})/(2(K-1))$ be the regret per draw with $\{\nu_{i,j}\}$ and

$$C_{i_1}^*(\{\nu_{i,j}\}) := \inf_{\{q_{i,j}\}_{i>j} \in \mathcal{R}_{i_1}(\{\nu_{i,j}\})} \sum_{(i,j)\in\mathcal{P}_{i\neq j}} \hat{r}_{i,j} q_{i,j} \,,$$

and let the (possibly non-unique) set of optimal solutions be

$$\mathcal{R}_{i_1}^*(\{\nu_{i,j}\}) := \Bigg\{ \{q_{i,j}\}_{i>j} \in \mathcal{R}_{i_1}(\{\nu_{i,j}\}) : \sum_{(i,j)\in\mathcal{P}_{i\neq j}} \hat{r}_{i,j} q_{i,j} = C_{i_1}^*(\{\nu_{i,j}\}) \Bigg\}.$$

The value $C_{i_1}^*(\{\mu_{i,j}\}) \log T$ is the possible minimum regret for exploration to make sure that the arm i_1 is in $[C]$. Using Lemma 1 yields the following regret lower bound.

Theorem 2. *The regret of a strongly consistent algorithm is lower bounded as:*

$$\mathbb{E}[R(T)] \geq \min_{i_1 \in [C]} C_{i_1}^*(\{\mu_{i,j}\}) \log T - o(\log T).$$

The proof of Theorem 2 is in Appendix E.

3.1. Comparison with the Consistency in Condorcet dueling bandits

A dueling bandit algorithm is strongly consistent in the sense of Condorcet if it has subpolynomial regret for any $M \in \mathcal{M}_{\mathrm{Cond}}$, where $\mathcal{M}_{\mathrm{Cond}}$ is the set of preference matrices in which the Condorcet winner (i.e., the Copeland winner i_1 of $L_{i_1} = 0$) exists (Komiyama et al., 2015a). Although the definitions of the regret in the two dueling bandit problems are slightly different, they are the same in that drawing pairs that include non-Copeland winners increases regret, and thus a subpolynomial regret in the sense of the Condorcet dueling bandit problem is consistent with the one of the Copeland dueling bandit problem. Therefore, a strongly consistent Copeland dueling bandit algorithm is also strongly consistent in the sense of the Condorcet dueling bandit problem since $\mathcal{M}_{\mathrm{Cop}} \supset \mathcal{M}_{\mathrm{Cond}}$. The converse is not necessarily true: when we run a Condorcet dueling bandit algorithm with a preference matrix without a Condorcet winner, it can fail to identify a Copeland winner.

An example in which the two consistencies make a difference is in Table 2. RMED2FH (Komiyama et al., 2015a), an optimal algorithm for solving the Condorcet dueling bandit problem, may not be consistent in the sense of Copeland; RMED2FH draws pairs $(2,3)$, $(2,4)$, and $(3,4)$ to prove that each of $2, 3$, and 4 is beaten by another arm, which implies that these arms are non-Condorcet. However, in the sense of Copeland, an algorithm must make sure that the superior of arm 1 is smaller than those of the other arms, and thus, it needs to compare arm 1 with the others for sufficiently many times.

4. Algorithms

In this section, we first introduce the CW-RMED algorithm, which is inspired by the DMED algorithm (Honda & Takemura, 2010) for solving the multi-armed bandit problem. We then derive an asymptotically optimal regret bound for CW-RMED. However, to the best of our knowledge, it is not known whether an optimization in the subroutine can be efficiently computed or not. To address this issue, we devise another algorithm called ECW-RMED, which is computationally efficient and has a regret bound that is close to optimal.

4.1. CW-RMED

Algorithm 1 is CW-RMED. At the beginning of each round $t = 1, 2, \ldots, T$, if there exists a pair $(i, j) \in \mathcal{P}_{i \neq j}$ that is not drawn $O(\sqrt{\log t})$ times or $\hat{\mu}_{i,j}(t)$ is very close to $1/2$, it immediately draws that pair. Otherwise, it enters the loop that sequentially draws each pair in L_C. After drawing each pair, it checks whether the current observation is sufficient or not. If the observation is enough to identify some $\hat{i}^*(t)$ as a Copeland winner, it exploits by adding $(\hat{i}^*(t), \hat{i}^*(t))$ into L_{NC}, the candidates of the pairs that will be drawn in the next loop. Otherwise, it draws the pairs with the number of observations below the minimum requirement for identifying $\hat{i}^*(t)$ as a winner with high confidence. Note that it considers a pair of the same arm (i, i) and pair of different arms $(i, j), i \neq j$, separately. Since a comparison with itself yields no information, drawing (i, i) is purely for exploitation.

The following theorem, whose proof is in Appendix H, states that the regret of CW-RMED is asymptotically optimal when we view the parameters of the preference matrix $\{\mu_{i,j}\}$ as constants. Therefore, it performs as well as any other strongly consistent algorithm for sufficiently large T.

Theorem 3. *Assume that* $\arg\min_{i_1 \in [C]} C^*_{i_1}(\{\mu_{i,j}\})$ *and* $\mathcal{R}^*_{i_1}(\{\mu_{i,j}\})$ *for each* $i_1 \in [C]$ *are unique. For any* $\alpha > 0$, $\beta > 0$, *the regret of CW-RMED is bounded as:*

$$\mathbb{E}[R(T)] \leq \min_{i_1 \in [C]} C^*_{i_1}(\{\mu_{i,j}\}) \log T + o(\log T)\,.$$

4.1.1. Computation of an optimal solution

Here, we discuss the computational aspects of CW-RMED. Checking (3) is relatively easy since we can sort $\{q_{i,j} d_{\mathrm{KL}}(\nu_{i,j}, 1/2)\}$ for each $(i_1, j) \in I_{i_1}$ or $(i_2, j) \in S_{i_2}$, and the constraint that matters is the top-c smallest of them for each size-c subset.

The difficult part is the computation of $\{q^*_{i,j}\} \in \mathcal{R}^*_{i_1}(\{\hat{\mu}_{i,j}(t)\})$ for each i_1, which can be formulated as a linear programming (LP). In the case of this paper the number of constraints of the LP is exponential in K and a naive use of an LP solver is sometimes very slow. It is well known that even if there are exponentially many constraints an LP can be solved by using the ellipsoid method (Khachiyan, 1980) in a polynomial time if there exists a polynomial-time oracle that (i) checks whether a point $\{q_{ij}\}$ is feasible or not and (ii) returns a hyperplane such that $\{q_{ij}\}$ and the feasible region are separated if $\{q_{ij}\}$ is infeasible. Such an oracle is easily constructed based on the

Table 2. A preference matrix of size 4×4. The ij-th element is $\mu_{i,j}$. The Copeland (Condorcet) winner is arm 1.

	1	2	3	4
1	0.5	0.6	0.6	0.6
2	0.4	0.5	0.9	0.1
3	0.4	0.1	0.5	0.9
4	0.4	0.9	0.1	0.5

Algorithm 1 CW-RMED and ECW-RMED Algorithms

Input: K arms, $\alpha > 0$, $\beta > 0$.
$L_C, L_R \leftarrow \mathcal{P}_{i \neq j}, L_N \leftarrow \emptyset$.
while $t \leq T$ **do**
  Draw all pairs such that $(i, j) \in \mathcal{P}_{i \neq j}$ if $N_{i,j}(t) < \alpha\sqrt{\log t}$ or $|\hat{\mu}_{i,j}(t) - 1/2| < \beta / \log\log t$. $t \leftarrow t + 1$ for each draw.
  for $p(t) = (l(t), m(t)) \in L_C$ in an arbitrarily fixed order **do**
    Draw arm pair $p(t)$.
    $L_{NC} \leftarrow \emptyset$.
    if

$$\{N_{i,j}(t)/\log t\}_{i \neq j} \in \mathcal{R}_{\hat{i}^*(t)}(\{\hat{\mu}_{i,j}(t)\}) \quad (3)$$

    for some $\hat{i}^*(t) \in \hat{\mathcal{C}}_{\mathrm{cop}}(\hat{\mu}_{i,j}(t))$ **then**
      Put $(\hat{i}^*(t), \hat{i}^*(t))$ into L_{NC}.
    else
      Compute some

$$\hat{i}^*(t) = \begin{cases} \arg\min_{i_1 \in \hat{\mathcal{C}}_{\mathrm{cop}}} C^*_{i_1}(\{\hat{\mu}_{i,j}(t)\}) & \text{(CW)} \\ \arg\min_{i_1 \in \hat{\mathcal{C}}_{\mathrm{cop}}} C^{\mathrm{E}*}_{i_1}(\{\hat{\mu}_{i,j}(t)\}) & \text{(ECW)} \end{cases}$$

$$\{q^*_{i,j}\} \in \begin{cases} \mathcal{R}^*_{\hat{i}^*(t)}(\hat{\mu}_{i,j}(t)) & \text{(CW)} \\ \mathcal{R}^{\mathrm{E}*}_{\hat{i}^*(t)}(\hat{\mu}_{i,j}(t)) & \text{(ECW)} \end{cases}$$

      (ties are broken arbitrarily) and put all pairs $(i, j) \in \mathcal{P}_{i \neq j}$ such that $q^*_{i,j} > N_{i,j}(t)/\log t$ into L_{NC}.
      Put $(\hat{i}^*(t), \hat{i}^*(t))$ into L_{NC}.
    end if
    $L_R \leftarrow L_R \setminus \{p(t)\}$.
    $L_N \leftarrow L_N \cup (i, j)$ (without a duplicate) for any $(i, j) \in L_{NC} \cap (\mathcal{P}_{i \neq j} \setminus L_R)$.
    $t \leftarrow t + 1$.
  end for
  $L_C, L_R \leftarrow L_N, L_N \leftarrow \emptyset$.
end while

sorting described above, and thus $\{q^*_{i,j}\} \in \mathcal{R}^*_{i_1}(\{\hat{\mu}_{i,j}(t)\})$ can be computed in a polynomial time. Although the ellipsoid method is practically very slow, a practical combinatorial algorithm is often derived later for many problems that are solvable by the ellipsoid method (see, e.g., Korte & Vygen 2007, Chapters 1–4 and 12). Thus the authors think that $\mathcal{R}^*_{i_1}(\{\hat{\mu}_{i,j}(t)\})$ can be computed practically. Still, in this paper, we consider a suboptimal solution because it runs not only in polynomial time but also in time almost the same as that of sorting, as described in Section 4.2.

4.2. ECW-RMED

In this section, we propose ECW-RMED (Algorithm 1). The difference between CW-RMED (Section 4.1) and ECW-RMED is the amount of exploration. For a candidate of Copeland winners i_1, it tries to make sure that neither $L_{i_1} \geq \min_i L_i + 1$ nor $L_{i_2} \leq \min_i L_i - 1$ for any $i_2 \neq i_1$ occurs, which implies that i_1 is a Copeland winner. Namely, for $i_1 \in \hat{\mathcal{C}}_{\mathrm{cop}}$, let

$$\mathcal{R}^{\mathrm{E}}_{i_1}(\{\nu_{i,j}\}) := \Bigg\{ \{q_{i,j}\}_{i>j} \in [0, 1/d_{\mathrm{KL}}(\nu_{i,j}, 1/2)]^{K(K-1)/2} : \forall j \in \hat{I}_{i_1}\, q_{i_1,j} = 1/d_{\mathrm{KL}}(\nu_{i_1,j}, 1/2), \tag{4}$$

$$\forall_{i_2 \neq i_1} \forall S \in \hat{\mathcal{S}}_{i_2}^{\setminus i_1, \hat{L}_{i_2} - \hat{L}_{i_1} + 1} \sum_{j \in S} q_{j,i_2} d_{\mathrm{KL}}(\nu_{j,i_2}, 1/2) \geq 1 \Bigg\}. \tag{5}$$

Note that the red lines are the differences from $\mathcal{R}_{i_1}(\cdot)$. Moreover, let

$$C^{\mathrm{E}*}_{i_1}(\{\nu_{i,j}\}) := \inf_{\{q_{i,j}\} \in \mathcal{R}_{i_1}(\{\nu_{i,j}\})} \sum_{(i,j) \in \mathcal{P}_{i \neq j}} \hat{r}_{i,j} q_{i,j}\,,$$

and let the (possibly non-unique) set of optimal solutions be

$$\mathcal{R}^{\mathrm{E}*}_{i_1}(\{\nu_{i,j}\}) := \Bigg\{ \{q_{i,j}\} \in \mathcal{R}^{\mathrm{E}}_{i_1}(\{\nu_{i,j}\}) : \sum_{(i,j) \in \mathcal{P}_{i \neq j}} \hat{r}_{i,j} q_{i,j} = C^*_{i_1}(\{\nu_{i,j}\}) \Bigg\}.$$

The following theorem, whose proof is in Appendix H, bounds the regret of ECW-RMED.

Theorem 4. *Assume that* $\arg\min_{i_1 \in [C]} C^{\mathrm{E}*}_i(\{\mu_{i,j}\})$ *and* $\mathcal{R}^{\mathrm{E}*}_{i_1}(\{\mu_{i,j}\})$ *for each* $i_1 \in [C]$ *are unique. For any* $\alpha > 0$, $\beta > 0$, *the regret of ECW-RMED is bounded as:*

$$\mathbb{E}[R(T)] \leq \min_{i_1 \in [C]} C^{\mathrm{E}*}_{i_1}(\{\mu_{i,j}\}) \log T + o(\log T)\,.$$

A quantitative discussion on the regret bounds of CW/ECW-RMED is found in Appendix C.

4.2.1. Efficient computation of ECW-RMED

In this section, we show an efficient method of finding $\{q_{i,j}\}_{i>j} \in \mathcal{R}^{\mathrm{E}*}_{i_1}(\{\hat{\mu}_{i,j}(t)\})$ for $i_1 \in \hat{\mathcal{C}}_{\mathrm{cop}}(\{\hat{\mu}_{i,j}(t)\})$. Since the inequality (5) is disjoint for each $i_2 \neq i_1$, solving it for each i_2 suffices. Let $\mathcal{S} := \hat{S}_{i_2} \setminus \{i_1\}$, $k := |\mathcal{S}| - (\hat{L}_{i_2} - \hat{L}_{i_1} + 1)$. Moreover, let $c_j := \hat{r}_{j,i_2}(\{\hat{\mu}_{i,j}(t)\})/d_{\mathrm{KL}}(\hat{\mu}_{j,i_2}(t), 1/2) \geq 0$ and $e_j := q_{j,i_2} d_{\mathrm{KL}}(\hat{\mu}_{j,i_2}(t), 1/2) \geq 0$. Accordingly, the regret minimization under (5) is reduced to the following linear optimization problem:

$$\begin{aligned} &\text{minimize} && \sum_{j \in \mathcal{S}} c_j e_j \\ &\text{subject to} && \forall_{S \subset \mathcal{S}: |S| = |\mathcal{S}| - k} \sum_{j \in S} e_j \geq 1. \end{aligned} \tag{6}$$

Here, $c_j \geq 0$ can be considered as a cost, and (6) is a cost minimization problem. In the following discussion we assume $|\mathcal{S}| > k > 0$; otherwise the optimization problem is trivial. The following theorem, whose proof is in Appendix F, states that an optimal solution of the problem is computed efficiently.

Theorem 5. *Let* $\sigma_1, \sigma_2, \ldots, \sigma_{|\mathcal{S}|} \in \mathcal{S}$ *be a permutation of* $\mathcal{S}$ *such that* $c_{\sigma_1} \leq c_{\sigma_2} \leq \cdots \leq c_{\sigma_{|\mathcal{S}|}}$. *There exists* $h > k$ *such that at least one optimal solution* $\{e^*_j\}$ *of* (6) *satisfies*

$$e^*_{\sigma_1} = e^*_{\sigma_2} = \cdots = e^*_{\sigma_h} = 1/(h-k), \quad e^*_{\sigma_{h+1}} = e^*_{\sigma_{h+2}} = \ldots e^*_{\sigma_{|\mathcal{S}|}} = 0. \tag{7}$$

Since we only have $|\mathcal{S}| - k \leq K$ candidates of h in (7), an optimal solution can be found by checking each of them.

4.3. Relation between CW-RMED and ECW-RMED

The following theorem, whose proof is in Appendix G, relates the optimal regret bound and the one of ECW-RMED.

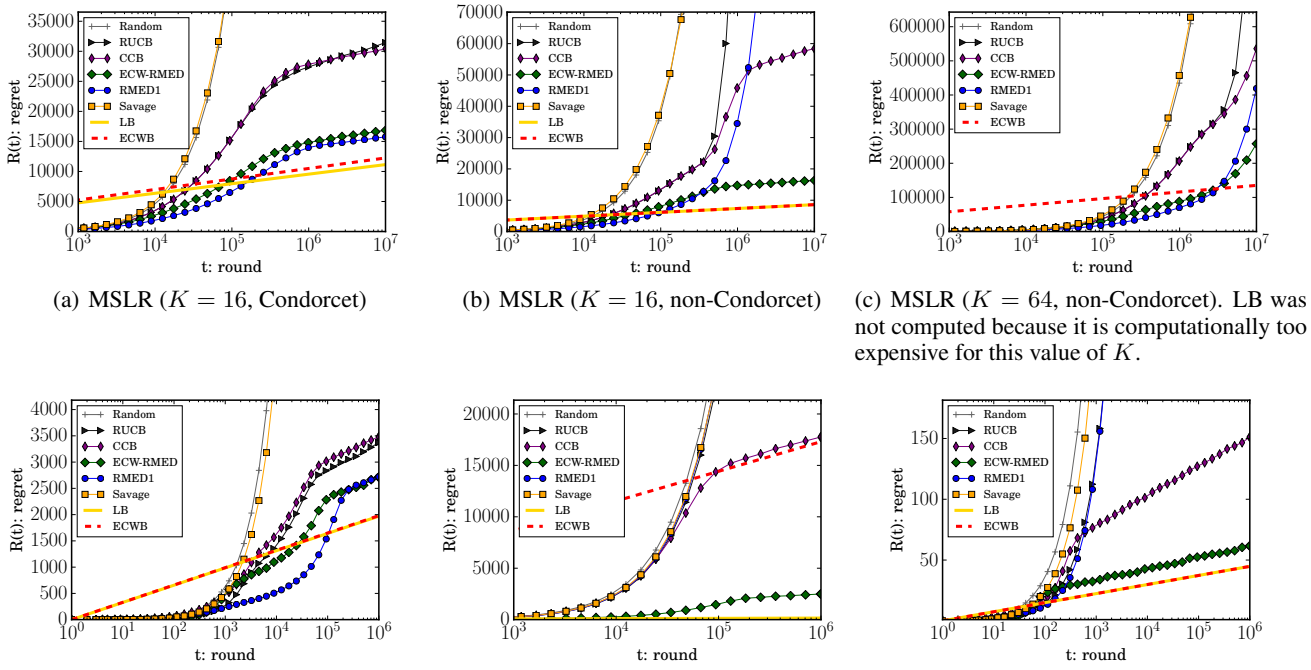
Theorem 6. (Optimality of ECW-RMED) *The following inequality always holds:*

$$C^{\mathrm{E}*}_{i_1}(\{\mu_{i,j}\}) \geq C^*_{i_1}(\{\mu_{i,j}\}). \tag{8}$$

Moreover, if $C \geq 2$, *the following equality holds:*

$$C^{\mathrm{E}*}_{i_1}(\{\mu_{i,j}\}) = C^*_{i_1}(\{\mu_{i,j}\}). \tag{9}$$

Inequality (8) states that the leading logarithmic constant of the bound on CW-RMED is always as good as that of ECW-RMED, which is natural since CW-RMED is asymptotically optimal as stated in Theorem 4. Still, (9) states that ECW-RMED has exactly the same constant when the Copeland winners are not unique.



(a) MSLR ($K = 16$, Condorcet)

(b) MSLR ($K = 16$, non-Condorcet)

(c) MSLR ($K = 64$, non-Condorcet). LB was not computed because it is computationally too expensive for this value of K.

(d) Sushi

(e) Gap; note that LB is very small (almost overlapping the bottom line).

(f) MultiSol

Figure 1. Regret-round semilog plots of algorithms. The regrets are averaged over 100 runs. LB and ECWB are the leading logarithmic terms of Theorems 2 and 4, respectively. One can see that ECWB is very close to LB on the MSLR $K = 16$ and the sushi datasets. We used the Gurobi LP solver for computing LB.

4.4. Comparison of ECW-RMED and CCB

In this section, we qualitatively discuss the improvement on the regret bound given by ECW-RMED.

Let $\Delta := \min_{(i,j)\in\mathcal{P}_{i\neq j}} |\mu_{i,j} - 1/2|$. Theorem 3 in Zoghi et al. (2015a) showed that CCB has an asymptotic regret bound[2] of

$$O\left(\frac{K(C + L_1 + 1)}{\Delta^2} \log T\right). \tag{10}$$

On the other hand, $\mathcal{R}^{\mathrm{E}*}_{i_1}(\{\mu_{i,j}\})$ includes

$$q_{i,j}/d_{\mathrm{KL}}(\mu_{i,j}, 1/2) = \begin{cases} 1 & \text{if } i = i_1, j \in I_{i_1}, \\ 1/(L_{i_2} - L_1 + 1) & \text{if } i = i_2, j \in S_{i_2} \setminus \{i_1\}, \\ 0 & \text{otherwise,} \end{cases}$$

which implies that ECW-RMED has a leading constant of

$$\min_{i_1 \in [C]} C^{\mathrm{E}*}_{i_1}(\{\mu_{i,j}\}) \leq \frac{1}{d_{\mathrm{KL}}(1/2 + \Delta, 1/2)} \sum_{i_2 \neq i_1} \left(1 + \frac{L_{i_2}}{L_{i_2} - L_1 + 1}\right). \tag{11}$$

[2]Here, we use a Δ that is a little bit looser than the one in the original bound of Zoghi et al. (2015a) for the sake of discussion. In Table 1, we used the value of the original regret bound of CCB.

This bound is expressed in terms of the KL divergence instead of $\Delta^2 \leq d_{\mathrm{KL}}(1/2 + \Delta, 1/2)/2$. Furthermore, taking the maximum of (11) over $\{L_{i_2}\}_{i_2 \neq i_1}$ with facts

$$\forall i_2 \neq i_1,\ L_1 \leq L_{i_2} \leq K, \qquad \sum_{i_2 \neq i_1} L_{i_2} \leq \frac{K^2}{2},$$

we see that

$$\min_{i_1} C^{\mathrm{E}*}_{i_1}(\{\mu_{i,j}\}) \leq \frac{K}{d_{\mathrm{KL}}(1/2 + \Delta, 1/2)} \left(\frac{L_1 + 3}{2} + \frac{L_1^2}{K}\right). \tag{12}$$

Therefore the regret of ECW-RMED can be bounded independent of C whereas (10) contains a $O(CK)$ term. Furthermore, the bound (12) is tight only in the case that L_{i_2} is close to L_1 for $O(K)$ arms i_2, which infrequently occurs in practice since $L_{i_2} \approx K/2$ on average. In fact, if $L_1 = o(K)$ and there exists $\rho \in (0, 1/2)$ such that $L_{i_2} \leq \rho K$ for at most $o(K)$ arms i_2 then we can bound (11) in the same way as (12) by

$$\min_{i_1} C^{\mathrm{E}*}_{i_1}(\{\mu_{i,j}\}) \leq \frac{2K + o(K)}{d_{\mathrm{KL}}(1/2 + \Delta, 1/2)},$$

which is independent of L_1.

The only drawback of our analysis is the assumption on the uniqueness of the optimal solution, which is not very

stringent. In our experiment, ECW-RMED performed well even when the optimal solution was not unique (MultiSol in Section 5).

4.5. On hyperparameters α and β

CW/ECW-RMED have two hyperparameters α and β. The hyperparameter α is necessary in both theoretical and practical point of views. It urges the draw of each pair for $o(\log t)$ times to assure the quality of the estimator $\hat{\mu}_{i,j}(t)$. On the other hand, we conjecture that the parameter β is a theoretical artifact. Technically, the hyperparameter β is required for bounding the regret when the quality of the estimation is low (i.e., inequality (32) in Appendix). A very small or zero β is practically sufficient: One can confirm that, setting $\beta = 0$ yields almost the same results as shown in Section 5.

5. Numerical Experiment

To evaluate the empirical performance of the proposed algorithms, we conducted computer simulations with the following datasets (preference matrices).

MSLR: We tested submatrices of a 136×136 preference matrix from Zoghi et al. (2015b), which is derived from the Microsoft Learning to Rank (MSLR) dataset (Microsoft Research, 2010; Qin et al., 2010) that consists of relevance information between queries and documents with more than 30K queries. Zoghi et al. (2015b) created a finite set of rankers, each of which corresponds to a ranking feature in the base dataset. The value $\mu_{i,j}$ is the probability that the ranker i beats ranker j based on the informational click model (Hofmann et al., 2013). We randomly chose subsets of rankers in our experiments and made sub preference matrices. We excluded cases with extremely small gaps such that $|\mu_{i,j} - 1/2| < 0.005$ for $K = 16$ or $|\mu_{i,j} - 1/2| < 0.0005$ for $K = 64$. Furthermore, we selected the submatrices in which the Condorcet winner exists (Figure 1(a)) and the Condorcet winner does not exist (Figures 1(b) and 1(c)).

Sushi: This dataset is based on the sushi preference dataset (Kamishima, 2003) that contains the preferences of $5,000$ Japanese users as regards to 100 types of sushi. We extracted 16 types of sushi and converted them into a preference matrix with $\mu_{i,j}$ corresponding to the ratio of users who prefer sushi i over j, which is shown in Table 3(a) in Appendix.

Gap is the preference matrix of Table 3(b) in Appendix. This matrix is a corner case in which $(\arg\min_{i_1} C^{\mathrm{E}*}_{i_1}(\{\mu_{i,j}\}))/(\arg\min_{i_1} C^{*}_{i^*}(\{\mu_{i,j}\})) > 100$.

MultiSol is the preference matrix of Table 3(c) in Appendix. This matrix is an example in which the optimality condition in Theorem 4 is violated.

Note that MLSR (Condorcet) and Sushi each have a Condorcet winner, whereas the others do not. The results with smaller preference matrices are shown in Appendix B.

Algorithms: We compared the following algorithms: Random is a uniformly random sampling among pairs. Copeland SAVAGE with $\delta = 1/T$ is the algorithm that is general enough to solve the Copeland dueling bandit problems and have $O(K^2 \log T)$ regret bounds. We did not include PBR and RankEI because the two algorithms are reported to be consistently outperformed by other algorithms (Zoghi et al., 2015a). RUCB (Zoghi et al., 2014) with $\alpha = 0.51$ and RMED1 (Komiyama et al., 2015a) are algorithms for solving Condorcet dueling bandit problems. These algorithms are not designed to find all instances of Copeland dueling bandit problems. The values of the hyperparameters of RMED1 are the same as in Komiyama et al. (2015a). CCB (Zoghi et al., 2015a) with $\alpha = 0.51$ and our ECW-RMED with $\alpha = 3.0$ and $\beta = 0.01$ are algorithms designed for the Copeland dueling bandit problems.

Results: Figure 1 plots the regrets of the algorithms. SAVAGE did not perform well for in any of the experiments. RMED1 performed best in MSLR (Condorcet). However, in datasets such as MSLR (non-Condorcet) and MultiSol where the Condorcet winner does not exist, it suffered a large regret. RUCB did not perform better than RMED1 and showed a similar tendency. These observations support the hypothesis that these algorithms are not capable of finding a Copeland winner. CCB performed similarly to RUCB in many datasets and outperformed RUCB for the datasets without a Condorcet winner. ECW-RMED significantly outperformed CCB and in all datasets, including Gap in which the uniqueness assumption of Theorem 4 is violated. In particular, in MSLR non-Condorcet dataset with $K = 16$, the regret of ECW-RMED was more than three times smaller than that of CCB. The slope of ECW-RMED in many of the datasets is close to ECWB when T is large, which is consistent with our analysis.

6. Conclusion

We studied the stochastic dueling bandit problem. The hardness of the problem of recommending Copeland winners was uncovered by deriving a lower bound of the regret. CW-RMED, an asymptotically optimal algorithm, was proposed. Moreover, ECW-RMED, a close-to-optimal algorithm, was proposed and an efficient computation method of it is given. ECW-RMED significantly outperformed the state-of-the-art algorithms in an experiment.

Acknowledgements

This work was supported in part by JSPS KAKENHI Grant Number 15J09850 and 16H00881.

References

Altman, Alon and Tennenholtz, Moshe. Axiomatic foundations for ranking systems. *J. Artif. Intell. Res. (JAIR)*, 31:473–495, 2008.

Bartók, Gábor, Foster, Dean P., Pál, Dávid, Rakhlin, Alexander, and Szepesvári, Csaba. Partial monitoring - classification, regret bounds, and algorithms. *Math. Oper. Res.*, 39(4):967–997, 2014.

Brochu, Eric, Brochu, Tyson, and de Freitas, Nando. A bayesian interactive optimization approach to procedural animation design. In *Proceedings of the 2010 Eurographics/ACM SIGGRAPH Symposium on Computer Animation, SCA 2010, Madrid, Spain, 2010*, pp. 103–112, 2010.

Busa-Fekete, Róbert, Szörényi, Balázs, Cheng, Weiwei, Weng, Paul, and Hüllermeier, Eyke. Top-k selection based on adaptive sampling of noisy preferences. In *ICML*, pp. 1094–1102, 2013.

Busa-Fekete, Róbert, Szörényi, Balázs, and Hüllermeier, Eyke. PAC rank elicitation through adaptive sampling of stochastic pairwise preferences. In *AAAI*, pp. 1701–1707, 2014.

Cover, Thomas M. and Thomas, Joy A. *Elements of information theory (2. ed.)*. Wiley, 2006. ISBN 978-0-471-24195-9.

Dudík, Miroslav, Hofmann, Katja, Schapire, Robert E., Slivkins, Aleksandrs, and Zoghi, Masrour. Contextual dueling bandits. In *COLT*, pp. 563–587, 2015.

Gajane, Pratik and Urvoy, Tanguy. Utility-based dueling bandits as a partial monitoring game. *CoRR*, abs/1507.02750v2, 2015. URL `http://arxiv.org/abs/1507.02750v2`.

Gemmis, Marco De, Iaquinta, Leo, Lops, Pasquale, Musto, Cataldo, Narducci, Fedelucio, and Semeraro, Giovanni. Preference learning in recommender systems. In *In Preference Learning (PL-09) ECML/PKDD-09 Workshop*, 2009.

Hofmann, Katja, Whiteson, Shimon, and de Rijke, Maarten. Fidelity, soundness, and efficiency of interleaved comparison methods. *Transactions on Information Systems*, 31(4):17:1–43, 2013.

Hogan, William W. Point-to-set maps in mathematical programming. *SIAM Review*, 15(3):591–603, 1973.

Honda, Junya and Takemura, Akimichi. An Asymptotically Optimal Bandit Algorithm for Bounded Support Models. In *COLT*, pp. 67–79, 2010.

Jamieson, Kevin G., Katariya, Sumeet, Deshpande, Atul, and Nowak, Robert D. Sparse dueling bandits. In *AISTATS*, 2015.

Kamishima, Toshihiro. Nantonac collaborative filtering: recommendation based on order responses. In *KDD*, pp. 583–588, 2003.

Khachiyan, L.G. Polynomial algorithms in linear programming. *USSR Computational Mathematics and Mathematical Physics*, 20(1):53 – 72, 1980. ISSN 0041-5553.

Komiyama, Junpei, Honda, Junya, Kashima, Hisashi, and Nakagawa, Hiroshi. Regret lower bound and optimal algorithm in dueling bandit problem. In *COLT*, pp. 1141–1154, 2015a.

Komiyama, Junpei, Honda, Junya, and Nakagawa, Hiroshi. Regret lower bound and optimal algorithm in finite stochastic partial monitoring. In *NIPS*, 2015b.

Korte, Bernhard and Vygen, Jens. *Combinatorial Optimization: Theory and Algorithms*. Springer Publishing Company, Incorporated, 4th edition, 2007. ISBN 3540718435, 9783540718437.

Lai, Tze Leung and Robbins, Herbert. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

Microsoft Research. Microsoft Learning to Rank Datasets, 2010. URL `http://research.microsoft.com/en-us/projects/mslr/`.

Qin, Tao, Liu, Tie-Yan, Xu, Jun, and Li, Hang. LETOR: A benchmark collection for research on learning to rank for information retrieval. *Inf. Retr.*, 13(4):346–374, 2010.

Urvoy, Tanguy, Clérot, Fabrice, Feraud, Raphaël, and Naamane, Sami. Generic exploration and k-armed voting bandits. In *ICML*, pp. 91–99, 2013.

Vanchinathan, Hastagiri P., Bartók, Gábor, and Krause, Andreas. Efficient partial monitoring with prior information. In *NIPS*, pp. 1691–1699, 2014.

Yue, Yisong and Joachims, Thorsten. Beat the mean bandit. In *ICML*, pp. 241–248, 2011.

Yue, Yisong, Broder, Josef, Kleinberg, Robert, and Joachims, Thorsten. The k-armed dueling bandits problem. In *COLT*, 2009.

Yue, Yisong, Broder, Josef, Kleinberg, Robert, and Joachims, Thorsten. The k-armed dueling bandits problem. *J. Comput. Syst. Sci.*, 78(5):1538–1556, 2012.

Zaidan, Omar and Callison-Burch, Chris. Crowdsourcing translation: Professional quality from non-professionals. In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA*, pp. 1220–1229, 2011.

Zoghi, Masrour, Whiteson, Shimon, Munos, Rémi, and de Rijke, Maarten. Relative upper confidence bound for the k-armed dueling bandit problem. In *ICML*, pp. 10–18, 2014.

Zoghi, Masrour, Karnin, Zohar Shay, Whiteson, Shimon, and de Rijke, Maarten. Copeland dueling bandits. In *NIPS*, 2015a.

Zoghi, Masrour, Whiteson, Shimon, and de Rijke, Maarten. Mergerucb: A method for large-scale online ranker evaluation. In *WSDM*, pp. 17–26, 2015b.

A. Preference Matrices in the Experiment

The following table shows the preference matrices that are used in the numerical experiment in Section 5.

Table 3. Preference matrices in the experiment. In each matrix, the ij element is $\mu_{i,j}$.

(a) Sushi. Rows are 1. mildly fatty tuna, 2. fatty tuna, 3. salmon, 4. tuna, 5. salmon roe, 6. sea bream, 7. sea eel, 8. scallop, 9. squid, 10. horse mackerel, 11. eel, 12. abalone, 13. mackerel, 14. squid feet, 15. Tori clam, 16. squilla, respectively.

0.5	0.512	0.622	0.655	0.698	0.726	0.711	0.708	0.749	0.8	0.741	0.783	0.847	0.817	0.854	0.868
0.488	0.5	0.602	0.683	0.652	0.776	0.663	0.683	0.738	0.709	0.786	0.802	0.83	0.85	0.871	0.873
0.378	0.398	0.5	0.528	0.554	0.533	0.534	0.591	0.573	0.593	0.661	0.705	0.734	0.672	0.787	0.822
0.345	0.317	0.472	0.5	0.553	0.619	0.566	0.641	0.675	0.687	0.665	0.696	0.803	0.823	0.796	0.844
0.302	0.348	0.446	0.447	0.5	0.513	0.524	0.518	0.608	0.538	0.643	0.61	0.695	0.672	0.681	0.775
0.274	0.224	0.467	0.381	0.487	0.5	0.513	0.559	0.575	0.621	0.591	0.701	0.702	0.787	0.829	0.811
0.289	0.337	0.466	0.434	0.476	0.487	0.5	0.559	0.553	0.613	0.564	0.607	0.703	0.735	0.736	0.801
0.292	0.317	0.409	0.359	0.482	0.441	0.441	0.5	0.556	0.527	0.562	0.58	0.668	0.805	0.777	0.767
0.251	0.262	0.427	0.325	0.392	0.425	0.447	0.444	0.5	0.512	0.548	0.542	0.612	0.786	0.71	0.685
0.2	0.291	0.407	0.313	0.462	0.379	0.387	0.473	0.488	0.5	0.543	0.579	0.613	0.718	0.685	0.747
0.259	0.214	0.339	0.335	0.357	0.409	0.436	0.438	0.452	0.457	0.5	0.564	0.625	0.618	0.702	0.684
0.217	0.198	0.295	0.304	0.39	0.299	0.393	0.42	0.458	0.421	0.436	0.5	0.542	0.644	0.7	0.733
0.153	0.17	0.266	0.197	0.305	0.298	0.297	0.332	0.388	0.387	0.375	0.458	0.5	0.577	0.607	0.596
0.183	0.15	0.328	0.177	0.328	0.213	0.265	0.195	0.214	0.282	0.382	0.356	0.423	0.5	0.578	0.637
0.146	0.129	0.213	0.204	0.319	0.171	0.264	0.223	0.29	0.315	0.298	0.3	0.393	0.422	0.5	0.586
0.132	0.127	0.178	0.156	0.225	0.189	0.199	0.233	0.315	0.253	0.316	0.267	0.404	0.363	0.414	0.5

(b) Gap

0.5	0.8	0.8	0.51	0.2
0.2	0.5	0.8	0.2	0.8
0.2	0.2	0.5	0.8	0.8
0.49	0.8	0.2	0.5	0.2
0.8	0.2	0.2	0.8	0.5

(c) MultiSol

0.5	0.2	0.8	0.8	0.8
0.8	0.5	0.2	0.8	0.8
0.2	0.8	0.5	0.8	0.8
0.2	0.2	0.2	0.5	0.6
0.2	0.2	0.2	0.4	0.5

(d) ArXiv

0.50	0.55	0.55	0.54	0.61	0.61
0.45	0.50	0.55	0.55	0.58	0.60
0.45	0.45	0.50	0.54	0.51	0.56
0.46	0.45	0.46	0.50	0.54	0.50
0.39	0.42	0.49	0.46	0.50	0.51
0.39	0.40	0.44	0.50	0.49	0.50

(e) MSLR (fixed, $K = 5$, Condorcet)

0.5	0.535	0.613	0.757	0.765
0.465	0.5	0.580	0.727	0.738
0.387	0.420	0.5	0.659	0.669
0.243	0.276	0.341	0.5	0.510
0.235	0.262	0.331	0.490	0.5

(f) MSLR Fixed ($K = 5$, non-Condorcet)

0.5	0.484	0.519	0.529	0.518
0.516	0.5	0.481	0.530	0.539
0.481	0.519	0.5	0.504	0.512
0.471	0.470	0.496	0.5	0.503
0.482	0.461	0.488	0.497	0.5

B. Additional Experiment

We conducted additional simulations with the following datasets.

ArXiv is a preference matrix based on the six retrieval functions in the full-text search engine of ArXiv.org (Yue & Joachims, 2011) shown in Table 3(d), where an order among arms exists. Although the fact $\mu_{4,6} = 1/2$ violates our assumption, the Copeland winner is arguably arm 1.

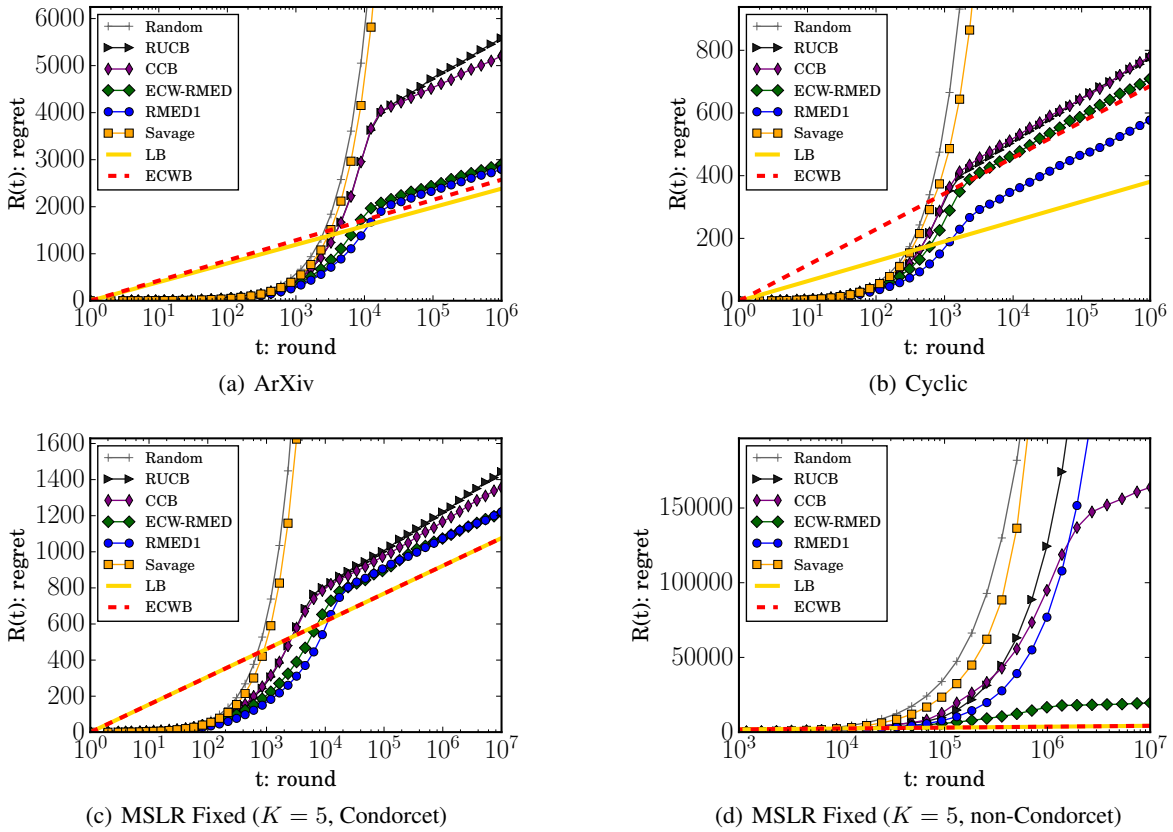
Cyclic is the preference matrix of Table 2.

MSLR Fixed are the two matrices of size 5×5 provided by Zoghi et al. (2015a) shown in Table 3(e) and 3(f). One matrix has a Condorcet winner, whereas the other does not. We include these matrices to compare our results with their ones.

The results of the simulations are shown in Figure 2.

C. Comparison of Regret Bounds

In this section, we clarify differences among the regret bounds of CW-RMED, ECW-RMED and CCB by calculating them in the cyclic preference matrix (Table 2). In the cyclic preference matrix, we have $C = 1, L_1 = 0$, and $\Delta = \min_{(i,j)\in\mathcal{P}_{i\neq j}} |\mu_{i,j} - 1/2| = 0.1$. Table 4 shows the regret bounds of the three algorithms. These bounds



(a) ArXiv (b) Cyclic

(c) MSLR Fixed ($K = 5$, Condorcet) (d) MSLR Fixed ($K = 5$, non-Condorcet)

Figure 2. Regret-round semilog plots of algorithms. The regrets are averaged over 100 runs. The algorithms, LB, and ECWB in the plots are the same as in the main text.

are calculated as follows. First, the regret bound of CW-RMED (inequality (2)) states that the risk of arm 1 being a non-Copeland winner is smaller than $\log T$: It requires $N_{1,2}(T), N_{1,3}(T), N_{1,4}(T) \geq (\log T)/(2d_{\mathrm{KL}}(0.6, 0.5))$ and $N_{2,3}(T), N_{3,4}(T), N_{4,2}(T) \geq (\log T)/(2d_{\mathrm{KL}}(0.9, 0.5))$. Second, the regret bound of ECW-RMED (inequality (5)) requires that (i) the arm 1 beats all other arms in I_1 and (ii) the other arms loses at least L_1 times: It requires $N_{1,2}(T), N_{1,3}(T), N_{1,4}(T) \geq (\log T)/d_{\mathrm{KL}}(0.6, 0.5)$. Finally, the regret bound of CCB ($(R(T)/\log T) = \frac{2K(C+L_1+1)}{\Delta^2} = 1600$) is much larger because it corresponds to the exploration for checking that (i) arm 1 wins against all other arms (CK pairs) and (ii) the other arms loses at least $L_1 + 1$ times ($(L_1 + 1)K$ pairs). Moreover, (iii) it may compare all pairs that are required to confirm (i)–(ii) for $2 \log T/\Delta^2$ times.

D. Facts

The following facts are frequently used in this paper. Fact 7 is a concentration inequality that bounds the tail probability on the empirical means. Fact 8 is used to bound the KL divergence from below. Fact 9 is later used in the proof of Lemma 11.

Table 4. Comparison of leading logarithmic constants of regret bounds in the cyclic dataset.

Optimal: CW-RMED	ECW-RMED	CCB
27.5	49.7	1600

Fact 7. (The Chernoff bound)
Let $X_1, \dots, X_n$ be i.i.d. binary random variables. Let $\hat{X} = \frac{1}{n}\sum_{i=1}^n X_i$ and $\mu = \mathbb{E}[\hat{X}]$. Then, for any $\epsilon > 0$,

$$\mathbb{P}(\hat{X} \geq \mu + \epsilon) \leq \exp\left(-d_{\mathrm{KL}}(\mu+\epsilon, \mu)n\right)$$

and

$$\mathbb{P}(\hat{X} \leq \mu - \epsilon) \leq \exp\left(-d_{\mathrm{KL}}(\mu-\epsilon, \mu)n\right).$$

Fact 8. (The Pinsker's inequality)
For $p, q \in (0,1)$, the KL divergence between two Bernoulli distributions is bounded as:

$$d_{\mathrm{KL}}(p, q) \geq 2(p-q)^2.$$

Fact 9. (Lemma 13 of Honda & Takemura 2010)
For any μ and μ_2 satisfying $0 < \mu_2 < \mu < 1$. Let $C_1(\mu, \mu_2) = (\mu - \mu_2)^2/(2\mu(1-\mu_2))$. Then, for any $\mu_3 \leq \mu_2$,

$$d_{\mathrm{KL}}(\mu_3, \mu) - d_{\mathrm{KL}}(\mu_3, \mu_2) \geq C_1(\mu, \mu_2) > 0.$$

E. Proofs on the Regret Lower Bound

In this section, we prove Lemma 1 and Theorem 2. In proofs, we frequently denote $\mathcal{A}, \mathcal{B}$ instead of $\mathcal{A} \cap \mathcal{B}$ for two events $\mathcal{A}$ and $\mathcal{B}$.

Proof of Lemma 1. Let $\delta > 0$ be arbitrary. For $i_1 \in [C]$, $i_2 \neq i_1$, $l \in \{\max\{0, L_1 - 1\}, \dots, L_2\}$, $I \in \mathcal{I}_{i_1}^{l+1-L_{i_1}}$, $S \in \mathcal{S}_{i_2}^{\setminus i_1, \max\{0, L_{i_2} - l - \mathbf{1}\{i_1 \in I\}\}}$, let

$$e^{\mathrm{Sum}}_{i_1,i_2,l,I,S}(T) := \sum_{(i,j)\in\mathcal{P}_{IS}} d_{\mathrm{KL}}(\mu_{i,j}, 1/2) N_{i,j}(T)$$

and

$$\mathcal{E}_{i_1,i_2,l,I,S}(T) := \{e^{\mathrm{Sum}}_{i_1,i_2,l,I,S}(T) \leq (1-\delta)\log T\}$$
$$\mathcal{A}(T) := \cap_{i_1} \cup_{i_2} \cup_l \cup_I \cup_S \mathcal{E}_{i_1,i_2,l,I,S}(T).$$

In the following we prove

$$\lim_{T\to\infty} \mathbb{P}[\mathcal{A}(T)] = 0,$$

which implies Lemma 1. Let

$$N_i^{\mathrm{sum}}(T) := \sum_{j\in[K]} N_{i,j}(T)$$

and

$$\mathcal{B}_i(T) := \left\{ N_i^{\mathrm{sum}}(T) = \max_{i'\in[K]} N_{i'}^{\mathrm{sum}}(T) \right\}.$$

Note that $\cup_{i\in[K]}\mathcal{B}_i(T)$ always occurs. Since $\mathcal{B}_{i_1}(T)$ implies $N_{i_1}^{\mathrm{sum}}(T) \geq T/K = \Omega(T)$, consistency requires $\mathbb{P}[\mathcal{B}_i(T)] = o(1)$ for each $i \notin [C]$ and thus

$$\mathbb{P}[\cup_{i\in[C]}\mathcal{B}_i(T)] = 1 - o(1). \tag{13}$$

Let $\epsilon_1 > 0$ be sufficiently small. Consider a modified preference matrix ${M'}_{i_1,i_2,l,I,S} := \{\mu'_{i,j}\} = \{\mu'_{i,j}(i_1, i_2, l, I, S)\}$ such that, for each pair (i,j) in $\mathcal{P}_{IS}$, if $\mu_{i,j} > 1/2$ then $\mu'_{i,j} < 1/2$ otherwise (i.e., if $\mu_{i,j} < 1/2$) $\mu'_{i,j} > 1/2$ such that

$$d_{\mathrm{KL}}(\mu_{i,j}, \mu'_{i,j}) = d_{\mathrm{KL}}(\mu_{i,j}, 1/2)(1+\epsilon_1). \tag{14}$$

Such a $\mu'_{i,j}$ for each pair (i,j) uniquely exists for sufficiently small ϵ_1. For each pair (i,j) that are not involved in $\mathcal{P}_{IS}$, we set $\mu'_{i,j} = \mu_{i,j}$. Let $\mathbb{E}' = \mathbb{E}'_{i_1,i_2,l,I,S}$, $\mathbb{P}' = \mathbb{P}'_{i_1,i_2,l,I,S}$ be the expectation and probability of the algorithm with respect to the modified preference matrix $M'_{i_1,i_2,l,I,S}$. Let $L'_i = \{j \in [K] : j \neq i, \mu'_{i,j} < 1/2\}$ be the number of arms that beat i in the modified game. In the modified game, arm i_1 is not a Copeland winner because $L'_{i_1} = l+1$ and $L'_{i_2} = l$. Let $\hat{X}^m_{i,j} \in \{0,1\}$ be the result of m-th draw of the pair (i,j),

$$\widehat{\mathrm{KL}}_{i,j}(n_{i,j}) = \sum_{m=1}^{n_{i,j}} \log\left(\frac{\hat{X}^m_{i,j}\mu_{i,j} + (1-\hat{X}^m_{i,j})(1-\mu_{i,j})}{\hat{X}^m_{i,j}\mu'_{i,j} + (1-\hat{X}^m_{i,j})(1-\mu'_{i,j})}\right),$$

and $\widehat{\mathrm{KL}}(\{n_{i,j}\}_{(i,j)\in\mathcal{P}_{IS}}) = \sum_{(i,j)\in\mathcal{P}_{IS}} \widehat{\mathrm{KL}}_{i,j}(n_{i,j})$. Let $\epsilon_2 > 0$ and

$$\mathcal{D}_{i_1,i_2,l,I,S}(T) := \left\{\widehat{\mathrm{KL}}(\{n_{i,j}\}_{(i,j)\in\mathcal{P}_{IS}}) < (1-\epsilon_2)\log T\right\}.$$

For any i_1, i_2, l, I, S,

$$\begin{aligned}
&\mathbb{P}[\mathcal{E}_{i_1,i_2,l,I,S}(T) \cap \mathcal{D}^c_{i_1,i_2,l,I,S}(T)] \\
&\leq \mathbb{P}[e^{\mathrm{Sum}}_{i_1,i_2,l,I,S}(T) \leq (1-\delta)\log T, \widehat{\mathrm{KL}}(\{n_{i,j}\}) > (1-\epsilon_2)\log T] \\
&= \mathbb{P}[\sum_{(i,j)\in\mathcal{P}_{IS}} d_{\mathrm{KL}}(\mu_{i,j},\mu'_{i,j})N_{i,j}(T) \leq (1-\delta)(1+\epsilon_1)\log T, \widehat{\mathrm{KL}}(\{n_{i,j}\}) > (1-\epsilon_2)\log T] \\
&\leq \mathbb{P}\left[\max_{\{n_{i,j}\}_{(i,j)\in\mathcal{P}_{IS}}\in\mathbb{N}^{|\mathcal{P}_{IS}|}, \sum_{(i,j)\in\mathcal{P}_{IS}} d_{\mathrm{KL}}(\mu_{i,j},\mu'_{i,j})n_{i,j} \leq (1-\delta)(1+\epsilon_1)\log T} \widehat{\mathrm{KL}}(\{n_{i,j}\}) > (1-\epsilon_2)\log T\right].
\end{aligned}$$

Note that,

$$\max_{1\leq n\leq N} \widehat{\mathrm{KL}}_{i,j}(n)$$

is the maximum sum of positive-mean random variables, and thus converges to its average. Namely,

$$\lim_{N\to\infty} \max_{1\leq n\leq N} \widehat{\mathrm{KL}}_{i,j}(n)/N = d_{\mathrm{KL}}(\mu_{i,j},\mu'_{i,j}) \qquad \text{a.s.}$$

and thus

$$\limsup_{T\to\infty} \frac{\max_{\{n_{i,j}\}_{(i,j)\in\mathcal{P}_{IS}}\in\mathbb{N}^{|\mathcal{P}_{IS}|}, \sum_{(i,j)\in\mathcal{P}_{IS}} d_{\mathrm{KL}}(\mu_{i,j},\mu'_{i,j})n_{i,j} \leq (1-\delta)(1+\epsilon_1)\log T} \widehat{\mathrm{KL}}(\{n_{i,j}\})}{\log T} \leq (1-\delta)(1+\epsilon_1) \qquad \text{a.s.}$$

Take $\epsilon_1 = \epsilon_1(\delta)$ and $\epsilon_2 = \epsilon_2(\delta)$ such that $(1-\delta)(1+\epsilon_1) < (1-\epsilon_2)$, and as a result

$$\lim_{T\to\infty} \mathbb{P}\left[\max_{\{n_{i,j}\}_{(i,j)\in\mathcal{P}_{IS}}\in\mathbb{N}^{|\mathcal{P}_{IS}|}, \sum_{(i,j)\in\mathcal{P}_{IS}} d_{\mathrm{KL}}(\mu_{i,j},\mu'_{i,j})n_{i,j} \leq (1-\delta)(1+\epsilon_1)\log T} \widehat{\mathrm{KL}}(\{n_{i,j}\}) > (1-\epsilon_2)\log T\right] = 0,$$

which leads to

$$\mathbb{P}[\mathcal{E}_{i_1,i_2,l,I,S}(T) \cap \mathcal{D}^c_{i_1,i_2,l,I,S}(T)] = o(1) \tag{15}$$

as a function of T.

Note that the consistency requires

$$\mathbb{P}'_{i_1,i_2,l,I,S}\{\mathcal{B}_{i_1}(T)\} = o(T^{a-1})$$

for any $a > 0$. Take $a < \epsilon_2$. For any i_1, i_2, l, I, S,

$$\begin{aligned}
&\mathbb{P}[\mathcal{B}_{i_1}(T) \cap \mathcal{D}_{i_1,i_2,l,I,S}(T)] \\
&= \sum_{T=\sum_{i=1}^K \sum_{j<i}^K n_{i,j}} \int_{\{N_{i,j}(T)=n_{i,j}\}} \mathbf{1}\{\mathcal{B}_{i_1}(T) \cap \mathcal{D}_{i_1,i_2,l,I,S}(T)\} \mathrm{e}^{\widehat{\mathrm{KL}}(\{n_{i,j}\}_{(i,j)\in\mathcal{P}_{IS}})} d\mathbb{P}'_{i_1,i_2,l,I,S} \\
&\leq T^{1-\epsilon_2}\mathbb{P}'_{i_1,i_2,l,I,S}[\mathcal{B}_{i_1}(T)] \leq o(T^{a-\epsilon_2}) = o(1).
\end{aligned} \tag{16}$$

We finally obtain

$$\begin{aligned}
\mathbb{P}[\mathcal{A}(T)] &= \mathbb{P}\left[\cap_{i_1\in[C]} \cup_{i_2} \cup_l \cup_I \cup_S \mathcal{E}_{i_1,i_2,l,I,S}(T)\right] \\
&\le \mathbb{P}\left[\cap_{i_1\in[C]} \cup_{i_2} \cup_l \cup_I \cup_S \{\mathcal{E}_{i_1,i_2,l,I,S}(T) \cap \mathcal{D}^c_{i_1,i_2,l,I,S}(T)\}\right] \\
&\quad + \mathbb{P}\left[\cap_{i_1\in[C]} \cup_{i_2} \cup_l \cup_I \cup_S \mathcal{D}_{i_1,i_2,l,I,S}(T)\right] \\
&= o(1) + \mathbb{P}\left[\cap_{i_1\in[C]} \cup_{i_2} \cup_l \cup_I \cup_S \mathcal{D}_{i_1,i_2,l,I,S}(T)\right] \qquad \text{(by union bound of (15) over } i_1,i_2,l,I,S).
\end{aligned}$$

Remember that $\cup_{i\in[C]}\mathcal{B}_i(T)$ occurs with probability $1-o(1)$ (inequality (13)). By using

$$\begin{aligned}
&\left\{\cap_{i_1\in[C]} \cup_{i_2} \cup_l \cup_I \cup_S \mathcal{D}_{i_1,i_2,l,I,S}(T)\right\} \cap \bigcup_{i\in[C]} \mathcal{B}_i(T) \\
&\subset \cup_{i_1\in[C]} \left\{\mathcal{B}_i(T) \cap (\cup_{i_2} \cup_l \cup_I \cup_S \mathcal{D}_{i_1,i_2,l,I,S}(T))\right\},
\end{aligned}$$

we have

$$\begin{aligned}
&\mathbb{P}\left[\cap_{i_1\in[C]} \cup_{i_2} \cup_l \cup_I \cup_S \mathcal{D}_{i_1,i_2,l,I,S}(T)\right] \\
&= \mathbb{P}\left[\cup_{i_1\in[C]} \left\{\mathcal{B}_i(T) \cap (\cup_{i_2} \cup_l \cup_I \cup_S \mathcal{D}_{i_1,i_2,l,I,S}(T))\right\}\right] + o(1) \qquad \text{(by (13))} \\
&= o(1) \qquad \text{(by union bound of (16) over } i_1,i_2,l,I,S).
\end{aligned}$$

In summary, $\mathbb{P}[\mathcal{A}(T)] = o(1)$ and thus the proof is completed.

□

Proof of Theorem 2. Assume that there exists $\delta > 0$ and a sequence $T_1 < T_2 < T_3 < \cdots$ such that for all s

$$\mathbb{E}[R(T_s)] < (1-\delta) \min_{i_1\in[C]} C^*_{i_1}(\{\mu_{i,j}\}) \log T_s \,,$$

that is, there exists $\mathcal{P}_{IS}(s)$ such that

$$\sum_{(i,j)\in\mathcal{P}_{IS}(s)} \frac{\mathbb{E}[N_{i,j}(T_s)]}{(1-\delta)\log T_s} r_{i,j} < \min_{i_1\in[C]} C^*_{i_1}(\{\mu_{i,j}\}) \,.$$

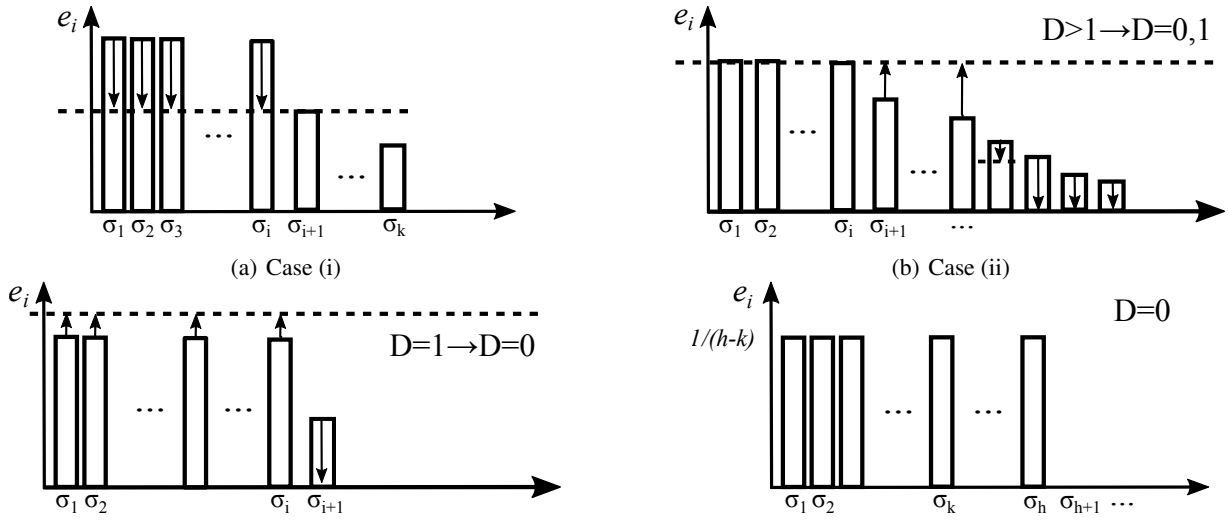
Let $i^* \in [C]$ be arbitrary and $\mathcal{S}$ be the closure of the space of preference matrices in which i_1 is not the Copeland winner, that is, $\mathcal{S} = \mathrm{cl}(\{\{\nu_{i,j}\}_{i>j} : i_1 \notin \hat{\mathcal{C}}_{\mathrm{cop}}(\{\nu_{i,j}\})\})$. From the definition of $C^*_{i_1}$, there exists $\{\nu_{i,j}(s)\} \in \mathcal{S}$ such that

$$\sum_{(i,j)\in\mathcal{P}_{IS}(s)} \frac{\mathbb{E}[N_{i,j}(T_s)]}{(1-\delta)\log T_s} d_{\mathrm{KL}}(\mu_{i,j}, \nu_{i,j}(s)) < 1 \,.$$

Since $\mathcal{S}$ is compact, there exists a subsequence $s_0 < s_1 < \cdots$ such that $\lim_{u\to\infty}\{\nu_{i,j}(s_u)\} = \{\nu'_{i,j}\}$ for some $\{\nu'_{i,j}\} \in \mathcal{S}$. Therefore from the lower semicontinuity of the divergence we obtain

$$\begin{aligned}
1 &\ge \liminf_{u\to\infty} \sum_{(i,j)\in\mathcal{P}_{IS}(s_u)} \frac{\mathbb{E}[N_{i,j}(T_{s_u})]}{(1-\delta)\log T_{s_u}} d_{\mathrm{KL}}(\mu_{i,j}, \nu_{i,j}(s_u)) \\
&\ge \liminf_{u\to\infty} \sum_{(i,j)\in\mathcal{P}_{IS}(s_u)} \frac{\mathbb{E}[N_{i,j}(T_{s_u})]}{(1-\delta)\log T_{s_u}} d_{\mathrm{KL}}(\mu_{i,j}, \nu'_{i,j}) \,,
\end{aligned}$$

which contradicts Lemma 1. □



(a) Case (i) (b) Case (ii) (c) Case (iii) (d) An optimal solution with $D = 0$.

Figure 3. On the solution of the optimization problem of (6). Figure (a)–(c) illustrate the operations that convert an optimal solution into another one with a smaller value of D. Figure (d) illustrates an optimal solution such that $D = 0$.

F. Proof on an Efficient Computation of ECW-RMED

Proof of Theorem 5. First, we show that there exists a optimal solution of (6) such that

$$e'_{\sigma_1} \geq e'_{\sigma_2} \geq \cdots \geq e'_{\sigma_{|\mathcal{S}|}} \tag{17}$$

for the following reason; let $\{e''_j\}$ be an arbitrary optimal solution. If there exists a pair $i < j$ such that $e''_{\sigma_i} > e''_{\sigma_j}$, swapping the values of e''_{σ_i} for e''_{σ_j} does not increase the objective value since $c_{\sigma_i} \leq c_{\sigma_j}$, and recursively applying this swap operation yields another optimal solution $\{e'_j\}$ such that (17) holds. The constraint in (6) for $\{e'_j\}$ satisfying (17) is equivalent to

$$\sum_{o \in k+1,\ldots,|\mathcal{S}|} e'_{\sigma_o} \geq 1. \tag{18}$$

Let the number of gaps be $D := \sum_{i=1}^{|\mathcal{S}|-1} \mathbf{1}\{e'_{\sigma_i} > e'_{\sigma_{i+1}}, e'_{\sigma_{i+1}} > 0\}$. In the following, we show that if $D > 0$ there exists another optimal solution with a smaller value of D. Let $D > 0$ and $i < |\mathcal{S}|$ be the smallest index such that $e'_{\sigma_i} > e'_{\sigma_{i+1}} > 0$. (i) if $i \leq k$, replacing $e'_{\sigma_1}, \ldots, e'_{\sigma_i}$ with $e'_{\sigma_{i+1}}$ does not increase the objective since each c_{σ_i} is non-negative. This operation yields another optimal solution that satisfies (17) and (18) with a smaller value of D, which is illustrated in Figure 3(a). (ii) If $i > k$ and $D \geq 2$, let $S = \sum_{j=i+1}^{|\mathcal{S}|} e'_{\sigma_j}$. Then,

$$e''_{\sigma_j} = \begin{cases} e'_{\sigma_i} & (\text{if } (j-i)e'_{\sigma_i} \geq S) \\ e'_{\sigma_i}(j-i) - S & (\text{if } (j-i+1)e'_{\sigma_i} \geq S > (j-i)e'_{\sigma_i}) \\ 0 & (\text{otherwise}) \end{cases}$$

has equal or smaller value of the objective since c_{σ_j} is non-decreasing in j. Therefore, $\{e''_j\}$ is an optimal solution with $D \leq 1$ such that (17) and (18) hold, which is illustrated in Figure 3(b). (iii) If $i > k$ and $D = 1$, then $\sum_{j=1}^{i} c_{\sigma_j} = (i-k)c_{\sigma_{i+1}}$ always hold. Otherwise, for sufficiently small $\delta > 0$ either of (iii-a) increasing $e'_{\sigma_1}, \ldots, e'_{\sigma_i}$ by δ and decreasing σ_{i+1} by $(i-k)\delta$ or (iii-b) decreasing $e'_{\sigma_1}, \ldots, e'_{\sigma_i}$ by δ and increasing σ_{i+1} by $(i-k)\delta$ must decrease the objective, which contradicts the assumption that $\{e'_j\}$ is optimal. Therefore, $\sum_{j=1}^{i} c_{\sigma_j} = (i-k)c_{\sigma_{i+1}}$. Then,

$$e''_{\sigma_j} = \begin{cases} e'_{\sigma_i} + e'_{\sigma_{i+1}}/(i-k) & (\text{if } j \leq i) \\ 0 & (\text{otherwise}) \end{cases}$$

has the same value of the objective function, and it satisfies (17) and (18). Therefore, $\{e''_j\}$ is an optimal solution with $D = 0$, which is illustrated in Figure 3(c). In summary, if $D > 0$, one can apply one of the operations (i)–(iii) that yields a modified optimal solution with a smaller value of D. Applying these operations yields the desired solution $\{e^*_j\}$ with $D = 0$, which is illustrated in Figure 3(d). □

G. Proof of Theorem 6

Proof of Theorem 6. First, one can check that (5) for each $i_2 \neq i_1$ is equivalent to the constraints of (2) for $l = L_1 - 1$. Second, Equation (4) implies that the constraints of (2) for all $i_2 \neq i_1, l \geq L_1$. Combining these two facts, we conclude that $\{q_{i,j}\} \in \mathcal{R}^{\mathrm{E}}_{i_1}(\{\mu_{i,j}\})$ implies $\{q_{i,j}\} \in \mathcal{R}_{i_1}(\{\mu_{i,j}\})$, and thus (8) is proven.

Moreover, to derive (9), it suffices to show $\mathcal{R}_{i_1}(\{\mu_{i,j}\}) = \mathcal{R}^{\mathrm{E}}_{i_1}(\{\mu_{i,j}\})$ for any preference matrix $\{\mu_{i,j}\}$ in which two or more Copeland winners exists. In that case, $L_1 = L_2$, and l in (2) runs for $\{L_1 - 1, L_1\}$. One can check that (4) is equivalent to the constraints of (2) for $l = L_1$. Since (5) for each $i_2 \neq i_1$ is equivalent to the constraints of (2) for $l = L_1 - 1$, the constraints of $\mathcal{R}_{i_1}(\{\mu_{i,j}\})$ and $\mathcal{R}^{\mathrm{E}}_{i_1}(\{\mu_{i,j}\})$ are equivalent.

□

H. Proofs of Theorems 3 and 4

In this section, we provide full proofs of Theorems 3 and 4. We define the following events that are important in bounding regret. Let

$$\mathcal{X}_{i,j}(t) := \left\{ \{N_{i,j}(t) < \alpha\sqrt{\log t}\} \cup \{|\hat{\mu}_{i,j}(t) - 1/2| < \beta/\log\log t\} \right\}$$

and $\mathcal{X}'_{i,j}(t)$ be the event that $\mathcal{X}_{i,j}(t)$ and pair (i,j) is drawn. Let $\mathcal{Y}_{i,j}(t)$ be the event that pair (i,j) is added into L_N. Note that $\cap_{(i',j')\in\mathcal{P}_{i\neq j}} \mathcal{X}^c_{i',j'}(t)$ implies the algorithm reaches Line 5 in Algorithm 1, and $\mathcal{Y}_{i,j}(t)$ implies $\cap_{(i',j')\in\mathcal{P}_{i\neq j}} \mathcal{X}^c_{i',j'}(t)$. Moreover, let

$$\mathcal{Z}_\delta(t) = \cap_{(i,j)\in\mathcal{P}_{i\neq j}} \{|\hat{\mu}_{i,j}(t) - \mu_{i,j}| < \delta\}.$$

In the following, we first show some lemmas, and then bounds the regret. The proofs of the lemmas are in the following sections of this appendix.

Lemma 10. (Case that arms are immediately drawn) *For CW/ECW-RMED, the following inequality holds:*

$$\sum_{t=1}^{T} \mathbb{P}[\mathcal{X}'_{i,j}(t)] \leq o(\log T).$$

Lemma 11. (Case that Copeland winner is not properly estimated) *For CW/ECW-RMED, for any $i_2 \in [K] \setminus [C]$ the following inequality holds:*

$$\sum_{t=1}^{T} \mathbb{P}\left[\bigcap_{(i',j')\in\mathcal{P}_{i\neq j}} \mathcal{X}^c_{i',j'}(t), \hat{i}^*(t) = i_2 \right] = O(1). \tag{19}$$

Lemma 12. (The continuity of the optimal solution) *Let the algorithm be CW-RMED. For $(i,j) \in \mathcal{P}_{i\neq j}$, let $R^*_{i,j}$ be the ij-th component of the unique element of $\mathcal{R}^*_{i^*}(\{\mu_{i,j}\})$ such that $i^* = \arg\min_{i_1\in[C]} C^*_{i_1}(\{\mu_{i,j}\})$. There exists $\epsilon(\delta)$ such that $\epsilon \to 0$ as $\delta \to +0$, and for any $(i,j) \in \mathcal{P}_{i\neq j}$,*

$$\sum_{t=1}^{T} \mathbf{1}[\mathcal{Y}_{i,j}(t), \mathcal{Z}_\delta(t)] \leq (1+\epsilon(\delta))R^*_{i,j}\log T + 1.$$

Let the algorithm be ECW-RMED. We can define $R^{\mathrm{E}*}_{i,j}$ in the same way and

$$\sum_{t=1}^{T} \mathbf{1}[\mathcal{Y}_{i,j}(t), \mathcal{Z}_\delta(t)] \leq (1+\epsilon(\delta))R^{\mathrm{E}*}_{i,j}\log T + 1.$$

Lemma 13. (The regret when the solution quality is low) *For CW/ECW-RMED, the following inequality holds:*

$$\sum_{t=1}^{T} \mathbb{P}\left[\bigcap_{(i',j') \in \mathcal{P}_{i \neq j}} \mathcal{X}^c_{i',j'}(t), \mathcal{Y}_{i,j}(t), \mathcal{Z}^c_\delta(t) \right] = o(\log T).$$

Note that the regret per round satisfies $r_{i,j} \leq 1$. For each pair (i,j) to be drawn, it either (i) satisfies $\mathcal{X}_{i,j}(t)$ (only for pair $i \neq j$), (ii) was put into L_N in the previous loop, or (iii) is in the first loop of L_C (only for pair $i \neq j$). By using this, the regret is bounded as

$$\begin{aligned}
R(T) &= \sum_{(i,j) \in \mathcal{P}_{\mathrm{all}}} r_{i,j} \sum_{t=1}^{T} \mathbf{1}[p(t) = (i,j)] \\
&\leq \sum_{(i,j) \in \mathcal{P}_{i \neq j}} \sum_{t=1}^{T} \mathbf{1}[\mathcal{X}'_{i,j}(t)] + \sum_{(i,j) \in \mathcal{P}_{\mathrm{all}}} r_{i,j} \sum_{t=1}^{T} \mathbf{1}[\cap_{(i',j') \in \mathcal{P}_{i \neq j}} \mathcal{X}^c_{i',j'}(t), \mathcal{Y}_{i,j}(t)] + \sum_{(i,j) \in \mathcal{P}_{i \neq j}} 1 \\
&\leq \sum_{(i,j) \in \mathcal{P}_{i \neq j}} \sum_{t=1}^{T} \mathbf{1}[\mathcal{X}'_{i,j}(t)] + \sum_{(i,i): i \in [K] \setminus [C]} \sum_{t=1}^{T} \mathbf{1}[\cap_{(i',j') \in \mathcal{P}_{i \neq j}} \mathcal{X}^c_{i',j'}(t), \mathcal{Y}_{i,i}(t)] + \sum_{(i,j) \in \mathcal{P}_{i \neq j}} r_{i,j} \sum_{t=1}^{T} \mathbf{1}[\mathcal{Y}_{i,j}(t), \mathcal{Z}_\delta(t)] \\
&\quad + \sum_{(i,j) \in \mathcal{P}_{i \neq j}} \sum_{t=1}^{T} (\mathbf{1}[\cap_{(i',j') \in \mathcal{P}_{i \neq j}} \mathcal{X}^c_{i',j'}(t), \mathcal{Y}_{i,j}(t), \mathcal{Z}^c_\delta(t)]) + K^2 \qquad (20)
\end{aligned}$$

In the following, we bound each term in (20) in expectation. First,

$$\sum_{(i,j) \in \mathcal{P}_{i \neq j}} \sum_{t=1}^{T} \mathbb{P}[\mathcal{X}'_{i,j}(t)] = o(\log T) \qquad (21)$$

follows from Lemma 10. Second,

$$\begin{aligned}
&\sum_{(i,i): i \in [K] \setminus [C]} \sum_{t=1}^{T} \mathbb{P}[\cap_{(i',j') \in \mathcal{P}_{i \neq j}} \mathcal{X}^c_{i',j'}(t), \mathcal{Y}_{i,i}(t)] \\
&\leq \sum_{(i,i): i \in [K] \setminus [C]} \sum_{t=1}^{T} \mathbb{P}[\cap_{(i',j') \in \mathcal{P}_{i \neq j}} \mathcal{X}^c_{i',j'}(t), \hat{i}^*(t) = i] \qquad \text{(by the fact that } \mathcal{Y}_{i,i}(t) \text{ implies } \hat{i}^*(t) = i) \\
&= O(1) \qquad \text{(by the union bound of Lemma 11 over } [K] \setminus [C]). \qquad (22)
\end{aligned}$$

Third, let the algorithm be CW-RMED. From Lemma 12,

$$\begin{aligned}
\sum_{(i,j) \in \mathcal{P}_{i \neq j}} r_{i,j} \sum_{t=1}^{T} \mathbf{1}[\mathcal{Y}_{i,j}(t), \mathcal{Z}_\delta(t)] &\leq (1 + \epsilon(\delta)) \sum_{(i,j) \in \mathcal{P}_{i \neq j}} r_{i,j} \left(R^*_{i,j} \log T + 1 \right) \\
&\leq (1 + \epsilon(\delta)) (\min_{i_1} C^*_{i_1}(\{\mu_{i,j}\}) \log T + K^2 \qquad (23)
\end{aligned}$$

The same arguments yields the following bound for ECW-RMED:

$$\sum_{(i,j) \in \mathcal{P}_{i \neq j}} r_{i,j} \sum_{t=1}^{T} \mathbf{1}[\mathcal{Y}_{i,j}(t), \mathcal{Z}_\delta(t)] \leq (1 + \epsilon(\delta)) (\min_{i_1} C^{\mathrm{E}*}_{i_1}(\{\mu_{i,j}\}) \log T + K^2.$$

Finally,

$$\sum_{(i,j) \in \mathcal{P}_{i \neq j}} \sum_{t=1}^{T} \mathbb{P}\left[\bigcap_{(i',j') \in \mathcal{P}_{i \neq j}} \mathcal{X}^c_{i',j'}(t), \mathcal{Y}_{i,j}(t), \mathcal{Z}^c_\delta(t) \right] = o(\log T) \qquad (24)$$

follows from Lemma 13.

Combining (20), (21), (22), (23), (24) completes the proof.

I. Proof of Lemma 10

Proof of Lemma 10. We have,

$$\sum_{t=1}^{T}\mathbb{P}[\mathcal{X}'_{i,j}(t)] = \sum_{n=1}^{T}\mathbb{P}\left[\bigcup_{t=n}^{T}\left\{N_{i,j}(t) < \alpha\sqrt{\log t} \cup |\hat{\mu}_{i,j}(t) - 1/2| < \beta/\log\log t, N_{i,j}(t) = n\right\}\right]$$
$$\leq \alpha\sqrt{\log T} + \sum_{n=1}^{T}\mathbb{P}\left[\bigcup_{t=n}^{T}\left\{|\hat{\mu}_{i,j}(t) - 1/2| < \beta/\log\log t, N_{i,j}(t) \geq \alpha\sqrt{\log T}, N_{i,j}(t) = n\right\}\right]. \quad (25)$$

Let $F(T) = \log\log(\alpha\sqrt{\log T})$. By using

$$|\hat{\mu}_{i,j}(t) - 1/2| \geq |\mu_{i,j} - 1/2| - |\hat{\mu}_{i,j}(t) - \mu_{i,j}|$$

we have

$$\sum_{n=1}^{T}\mathbb{P}\left[\bigcup_{t=n}^{T}\left\{|\hat{\mu}_{i,j}(t) - 1/2| < \beta/\log\log t, N_{i,j}(t) \geq \alpha\sqrt{\log T}, N_{i,j}(t) = n\right\}\right]$$
$$\leq \sum_{n=1}^{T}\mathbb{P}[|\hat{\mu}^n_{i,j} - \mu_{i,j}| > \beta/F(T)] + \sum_{n=1}^{T}\mathbb{P}[|\mu_{i,j} - 1/2| < (2\beta)/(\log\log n)]$$
$$\leq \sum_{n=1}^{T}\mathbb{P}[|\hat{\mu}^n_{i,j} - \mu_{i,j}| > \beta/F(T)] + e^{e^{2\beta/|\mu_{i,j}-1/2|}}$$
$$\leq 2\sum_{n=1}^{\infty} e^{-2n(\beta/F(T))^2} + e^{e^{\beta/(2|\mu_{i,j}-1/2|)}} \text{ (by Chernoff bound and Pinsker's inequality)}$$
$$\leq O\left(\frac{F(T)^2}{\beta^2}\right) + e^{e^{\beta/(2|\mu_{i,j}-1/2|)}} = o(\log T). \quad (26)$$

Combining (25) and (26) completes the proof. □

J. Proof of Lemma 11

Proof of Lemma 11. Note that we can assume $\hat{\mu}_{i,j}(t) \neq 1/2$ from $\mathcal{X}^c_{i,j}(t)$. Let $i_1 \in [C]$ be arbitrary. Event $\{\hat{i}^*(t) = i_2\}$ implies that, there exists a set of pairs $\mathcal{P}_{IS}$ such that $l \in \{\max\{0, L_1 - 1\}, \ldots, L_{i_2}\}$, $I \in \mathcal{I}_{i_1}^{l+1-L_1}$, $S \in \mathcal{S}_{i_2}^{\setminus i_1, \max\{0, L_{i_2} - l - \mathbf{1}\{i_1 \in I\}\}}$ and $\mathcal{P}_{IS} = \{(i_1, j) : j \in I\} \cup \{(i_2, j) : j \in S\}$ and the signs of $\hat{\mu}_{i,j}(t) - 1/2$ and $\mu_{i,j} - 1/2$ are different. In other words,

$$\{\hat{i}^*(t) = i_2\} \subset \{\hat{i}^*(t) = i_2\} \cap \left\{\cup_l \cup_I \cup_S \cap_{(i,j)\in\mathcal{P}_{IS}} \{(\hat{\mu}_{i,j}(t) - 1/2)(\mu_{i,j} - 1/2) < 0\}\right\}.$$

In the following we are going to show

$$\sum_{t=1}^{T}\mathbb{P}\left[\cap_{(i',j')\in\mathcal{P}_{i\neq j}}\mathcal{X}^c_{i',j'}(t), \hat{i}^*(t) = i_2, \cap_{(i,j)\in\mathcal{P}_{IS}}\{(\hat{\mu}_{i,j}(t) - 1/2)(\mu_{i,j} - 1/2) < 0\}\right] = O(1) \quad (27)$$

for each l, I, S. Note that

$$\left\{\log t \geq \sum_{(i,j)\in\mathcal{P}_{IS}} N_{i,j}(t) d_{\mathrm{KL}}(\hat{\mu}_{i,j}(t), 1/2), \cap_{(i',j')\in\mathcal{P}_{i\neq j}}\mathcal{X}^c_{i',j'}(t), \hat{i}^*(t) = i_2\right\}$$

implies that $\{N_{i,j}(t)/\log t\} \notin \mathcal{R}_{i_2}(\{\hat{\mu}_{i,j}(t)\})$ and at least one of the pairs in $\mathcal{P}_{IS}$ is immediately put into L_{NC} to satisfy the constraints. Therefore, one of the arms in $\mathcal{P}_{IS}$ is drawn within K^2 rounds of $\{t : \cap_{(i',j')\in\mathcal{P}_{i\neq j}}\mathcal{X}^c_{i',j'}(t)\}$. By using this

fact, we have

$$\sum_t \mathbf{1}\left[\cap_{(i',j')\in\mathcal{P}_{i\neq j}}\mathcal{X}^c_{i',j'}(t), \hat{i}^*(t)=i_2, \cap_{(i,j)\in\mathcal{P}_{IS}}\{(\mu_{i,j}-1/2)(\hat{\mu}_{i,j}(t)-1/2)<0, N_{i,j}(t)=n_{i,j}\}\right] \leq \exp\left(\sum_{(i,j)\in\mathcal{P}_{IS}} n_{i,j} d_{\mathrm{KL}}(\hat{\mu}_{i,j}(t), 1/2)\right) + K^2.$$

Let $\hat{\mu}^n_{i,j}$ be the empirical estimate of $\mu_{i,j}$ with n draws. Letting $P_{i,j}(x_{i,j}) = \mathbb{P}[(\mu_{i,j}-1/2)(\hat{\mu}^{n_{i,j}}_{i,j}-1/2) \leq 0, d_{\mathrm{KL}}(\hat{\mu}^{n_{i,j}}_{i,j}, 1/2) \geq x_{i,j}]$, we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_t \mathbf{1}\left[\bigcap_{(i',j')\in\mathcal{P}_{i\neq j}}\mathcal{X}^c_{i',j'}(t), \hat{i}^*(t)=i_2, \bigcap_{(i,j)\in\mathcal{P}_{IS}}\{(\mu_{i,j}-1/2)(\hat{\mu}^{n_{i,j}}_{i,j}-1/2)<0,\ N_{i,j}(t)=n_{i,j}\}\right]\right] \\
&\leq \int_{\{x_{i,j}\}\in[0,\log 2]^{|\mathcal{P}_{IS}|}} \left(\exp\left(\sum_{(i,j)\in\mathcal{P}_{IS}} n_{i,j}x_{i,j}\right) + K^2\right) \prod_{(i,j)\in\mathcal{P}_{IS}} \mathrm{d}(-P_{i,j}(x_{i,j})) \\
&= K^2 \prod_{(i,j)\in\mathcal{P}_{IS}} P_{i,j}(0) + \prod_{(i,j)\in\mathcal{P}_{IS}} \int_{x_{i,j}\in[0,\log 2]} \mathrm{e}^{n_{i,j}x_{i,j}} \mathrm{d}(-P_{i,j}(x_{i,j})) \\
&= K^2 \prod_{(i,j)\in\mathcal{P}_{IS}} P_{i,j}(0) + \prod_{(i,j)\in\mathcal{P}_{IS}} \left([-\mathrm{e}^{n_{i,j}x_{i,j}}P_{i,j}(x_{i,j})]_0^{\log 2} + \int_{x_{i,j}\in[0,\log 2]} n_{i,j}\mathrm{e}^{n_{i,j}x_{i,j}}P_{i,j}(x_{i,j})\mathrm{d}x_{i,j}\right) \\
&\quad \text{(integration by parts)} \\
&\leq (1+K^2) \prod_{(i,j)\in\mathcal{P}_{IS}} P_{i,j}(0) + \prod_{(i,j)\in\mathcal{P}_{IS}} \int_{x_{i,j}\in[0,\log 2]} n_{i,j}\mathrm{e}^{n_{i,j}x_{i,j}}\mathrm{e}^{-n_{i,j}(x_{i,j}+C_1(\mu_{i,j},1/2))}\mathrm{d}x_{i,j} \\
&\quad \text{(by Chernoff bound and Fact 9, where } C_1(\mu,\mu_2)=(\mu-\mu_2)^2/(2\mu(1-\mu_2))) \\
&\leq (1+K^2) \prod_{(i,j)\in\mathcal{P}_{IS}} \mathrm{e}^{-n_{i,j}d_{\mathrm{KL}}(1/2,\mu_{i,j})} + \prod_{(i,j)\in\mathcal{P}_{IS}} \int_{x_{i,j}\in[0,\log 2]} n_{i,j}\mathrm{e}^{-n_{i,j}C_1(\mu_{i,j},1/2)}\mathrm{d}x_{i,j} \\
&= (1+K^2) \prod_{(i,j)\in\mathcal{P}_{IS}} \mathrm{e}^{-n_{i,j}d_{\mathrm{KL}}(1/2,\mu_{i,j})} + \prod_{(i,j)\in\mathcal{P}_{IS}} (\log 2) n_{i,j}\mathrm{e}^{-n_{i,j}C_1(\mu_{i,j},1/2)}. \qquad (28)
\end{aligned}$$

By summing (28) over $\{n_{i,j}\}$,

$$\begin{aligned}
&\sum_{t=1}^T \mathbb{P}\left[\bigcap_{(i',j')\in\mathcal{P}_{i\neq j}}\mathcal{X}^c_{i',j'}(t), \bigcap_{(i,j)\in\mathcal{P}_{IS}}\{(\mu_{i,j}-1/2)(\hat{\mu}^{n_{i,j}}_{i,j}-1/2)<0\}\right] \\
&\leq \sum\cdots\sum_{\{n_{i,j}\}\in\mathbb{N}^{|\mathcal{P}_{IS}|}} \left((1+K^2)\prod_{(i,j)\in\mathcal{P}_{IS}} \mathrm{e}^{-n_{i,j}d_{\mathrm{KL}}(1/2,\mu_{i,j})} + \prod_{(i,j)\in\mathcal{P}_{IS}} (\log 2) n_{i,j}\mathrm{e}^{-n_{i,j}C_1(\mu_{i,j},1/2)}\right) \\
&\leq (1+K^2) \prod_{(i,j)\in\mathcal{P}_{IS}} \frac{1}{\mathrm{e}^{d_{\mathrm{KL}}(1/2,\mu_{i,j})}-1} + (\log 2)^{|\mathcal{P}_{IS}|} \prod_{(i,j)\in\mathcal{P}_{IS}} \frac{\mathrm{e}^{C_1(\mu_{i,j},1/2)}}{(\mathrm{e}^{C_1(\mu_{i,j},1/2)}-1)^2}, \\
&= O(1)
\end{aligned}$$

where we used the fact that $\sum_{n=1}^\infty \mathrm{e}^{-nx} = 1/(\mathrm{e}^x+1)$ and $\sum_{n=1}^\infty n\mathrm{e}^{-nx} = \mathrm{e}^x/(\mathrm{e}^x+1)^2$. In summary, we showed (27). Taking a union bound over l, I, S yields (19). □

K. Proof of Lemma 12

Following Hogan (1973), we define the continuity of a point-to-set map $\Omega : X \to 2^Y$ between metric spaces X and Y as follows: (i) Ω is open at $x_0 \in X$ if $\{x^k\}$, $x^k \to x_0$, and $y_0 \in \Omega(x_0)$ imply the existence of an integer m and a sequence

$\{y^k\}$ such that $y^k \in \Omega(x^k)$ for $k \geq m$ and $y^k \to y_0$. (ii) Ω is closed at x_0 if $\{x^k\} \in X$, $x^k \to x_0$, $y^k \to y_0$ imply that $y_0 \in \Omega(x_0)$. Moreover, (iii) Ω is continuous at x_0 if it is closed and open at x_0.

Let a set of relaxed feasible solutions be

$$\mathcal{R}^{+1}_{i_1}(\{\nu_{i,j}\}) := \Bigg\{ \{q_{i,j}\}_{i>j} \in [0, 1/d_{\mathrm{KL}}(\nu_{i,j}, 1/2)+1]^{K(K-1)/2} : \forall_{i_2 \neq i_1}\, \forall l \in \{\max\{0, \hat{L}^{(1)} - 1\}, \dots, \hat{L}^{(2)}\}$$
$$\forall I \in \hat{\mathcal{I}}^{(l+1-\hat{L}^{(1)})}_{i_1}\, \forall S \in \hat{\mathcal{S}}^{\setminus i_1, \max\{0, \hat{L}_{i_2} - l - \mathbf{1}\{i_2 \in I\}\}}_{i_2} \sum_{(i,j) \in \mathcal{P}_{IS}} q_{i,j} d_{\mathrm{KL}}(\nu_{i,j}, 1/2) \geq 1 \Bigg\}.$$

Note that the red term is the difference from $\mathcal{R}_{i_1}(\cdot)$. This set of relaxed feasible solutions is introduced for the sake of inequality (29) that appears later. The optimal coefficient $C^{+1,*}_{i_1}(\{\nu_{i,j}\})$ and the set of the optimal solutions $\mathcal{R}^{+1,*}_{i_1}(\{\nu_{i,j}\})$ are defined in accordance with $\mathcal{R}^{+1}_{i_1}(\{\nu_{i,j}\})$, that is,

$$C^{+1,*}_{i_1}(\{\nu_{i,j}\}) := \inf_{\{q_{i,j}\}_{i>j} \in \mathcal{R}^{+1}_{i_1}(\{\nu_{i,j}\})} \sum_{(i,j) \in \mathcal{P}_{i \neq j}} \hat{r}_{i,j} q_{i,j}\,,$$

and

$$\mathcal{R}^{+1,*}_{i_1}(\{\nu_{i,j}\}) := \left\{ \{q_{i,j}\}_{i>j} \in \mathcal{R}^{+1}_{i_1}(\{\nu_{i,j}\}) : \sum_{(i,j) \in \mathcal{P}_{i \neq j}} \hat{r}_{i,j} q_{i,j} = C^{+1,*}_{i_1}(\{\nu_{i,j}\}) \right\}.$$

Let the norms on $\{\nu_{i,j}\}$ and $\{q_{i,j}\}$ be $|\{\nu_{i,j}\}| = \sum_{i,j} |\nu_{i,j}|$ and $|\{q_{i,j}\}| = \sum_{i,j} |q_{i,j}|$, respectively. In the following, we show the following lemma:

Lemma 14. (The continuity of the solution function) *The point-to-set map $\mathcal{R}^{+1,*}_{i_1}(\{\nu_{i,j}\}) : \mathcal{M}_{\mathrm{Cop}} \to 2^{[0,\infty)^{K(K-1)}}$ is continuous at $\{\nu_{i,j}\} = \{\mu_{i,j}\}$.*

The continuity and the uniqueness of the optimal solution function $\mathcal{R}^{+1,*}_{i_1}(\{\mu_{i,j}\})$ implies that all solutions of $\mathcal{R}^{+1,*}_{i_1}(\{\nu_{i,j}\})$ approach $\mathcal{R}^{+1,*}_{i_1}(\{\mu_{i,j}\})$ ($= \mathcal{R}^{*}_{i_1}(\{\mu_{i,j}\})$, unique) when $\{\nu_{i,j}\}$ is sufficiently close to $\{\mu_{i,j}\}$. To prove Lemma 14, we first restate the following three Lemmas of Hogan (1973):

Lemma 15. (Theorem 10 of Hogan 1973) *Let g be a set of real-valued functions on $X \times Y$, and $P(x) := \{y \in Y : g(x, y) \leq 0\}$ be a map of feasible solutions. If each component of g is continuous on $x_0 \times Y$, then P is closed at x_0.*

Lemma 16. (Theorem 12 of Hogan 1973) *If Y is convex and normed, if each component of g is continuous on $x_0 \times P(x_0)$ and convex in y for each fixed $x \in X$, and if there exists a y_0 such that $g(x_0, y_0) < 0$, then P is open at x_0.*

Lemma 17. (Corollary 8.1 of Hogan 1973) *Let $\Omega : X \to 2^Y$ be a point-to-set map and $M(x) := \{y \in \Omega(x) : \sup_{y' \in \Omega(x)} f(x, y') = f(x, y)\}$ be an optimal solution function of some real-valued function f on $X \times Y$. Suppose Ω is continuous at x_0, f is continuous on $x_0 \times \Omega(x_0)$, M is non-empty and uniformly compact near x_0, and $M(x_0)$ is unique. Then, M is continuous at x_0.*

Proof of Lemma 14. We first show the continuity of the feasible solution function $\mathcal{R}^{+1}_{i}(\{\nu_{i,j}\})$ at $\{\nu_{i,j}\} = \{\mu_{i,j}\}$. The continuity of each component of g as a function of $\{\nu_{i,j}\}, \{q_{i,j}\}$ follows from the continuity of the KL divergence, and thus, applying Lemma 15 for $P = \mathcal{R}^{+1}_{i_1}$, $x_0 = \{\mu_{i,j}\}$ and $g(\{\nu_{i,j}\}, \{q_{i,j}\}) = \{1 - \sum_{(i,j) \in \mathcal{P}_{IS}} q_{i,j} d_{\mathrm{KL}}(\nu_{i,j}, 1/2)\}_{i_2, l, I, S}$ yields the closedness of $\mathcal{R}^{+1}_{i_1}$ at $\{\mu_{i,j}\}$. Moreover, by (i) continuity of each component of g, (ii) linearity of each component of g as a function of $\{q_{i,j}\}$ for each $\{\nu_{i,j}\}$, and (iii) the fact that $\{q'_{i,j}\} := \{(1/d_{\mathrm{KL}}(\nu_{i,j}, 1/2)) + 1\}^{K(K-1)}$ satisfies

$$\sum_{(i,j) \in \mathcal{P}_{IS}} q'_{i,j} d_{\mathrm{KL}}(\nu_{i,j}, 1/2) > 1, \tag{29}$$

applying Lemma 16 to the same P, x_0, g and $y_0 = \{(1/d_{\mathrm{KL}}(\nu_{i,j}, 1/2)) + 1\}$ yields the openness of $\mathcal{R}^{+1}_{i_1}$ at $\{\mu_{i,j}\}$. The continuity of $\mathcal{R}^{+1}_{i_1}$ follows from its closedness and the openness.

Finally, by using the continuity of $\mathcal{R}^{+1}_{i_1}$ and $C^{+1,*}_{i}$, and uniform compactness and uniqueness of $\mathcal{R}^{+1,*}_{i_1}$ at $\{\mu_{i,j}\}$, applying Lemma 17 to $M = \mathcal{R}^{+1,*}_{i_1}$, $\Omega = \mathcal{R}^{+1}_{i_1}$, and $f = C^{*}_{i_1}$ yields the continuity of $\mathcal{R}^{+1,*}_{i_1}$ at $\{\mu_{i,j}\}$. □

Proof of Lemma 12. By using the continuity of $\mathcal{R}_{i_1}^{+1,*}(\{\nu_{i,j}\})$ (Lemma 14), $\mathcal{R}_{i_1}^{*}(\{\nu_{i,j}\}) \subset \mathcal{R}_{i_1}^{+1,*}(\{\nu_{i,j}\})$, and the uniqueness of $\arg\min_{i_1 \in [C]} C_{i_1}^{*}(\{\mu_{i,j}\})$ and $\mathcal{R}_{i_1}^{*}(\{\mu_{i,j}\})$, there exists $\epsilon(\delta)$ such that $\epsilon \to 0$ as $\delta \to +0$ and

$$\begin{aligned}\sum_{t=1}^{T} \mathbf{1}[\mathcal{Y}_{i,j}(t), \mathcal{Z}_\delta(t)] &\le \sum_{n=1}^{T} \mathbf{1}\left[\bigcup_{t=1}^{T} \{\mathcal{Y}_{i,j}(t), \mathcal{Z}_\delta(t), N_{i,j}(t) = n\}\right] \\ &\le \sum_{n=1}^{T} \mathbf{1}\left[\bigcup_{t=1}^{T} \left\{n/\log t \le (1+\epsilon(\delta))R_{i,j}^{*}\right\}\right] \\ &\le (1+\epsilon(\delta))R_{i,j}^{*}\log T + 1.\end{aligned}$$

The same arguments also applies to ECW-RMED. □

L. Proof of Lemma 13

Proof of Lemma 13. We have

$$\begin{aligned}&\sum_{t=1}^{T} \mathbb{P}\left[\bigcap_{(i',j')\in\mathcal{P}_{i\neq j}} \mathcal{X}_{i',j'}^{c}(t), \mathcal{Y}_{i,j}(t), \mathcal{Z}_\delta^{c}(t)\right] \\ &\le \sum_{t=1}^{T} \mathbb{P}\left[\bigcap_{(i',j')\in\mathcal{P}_{i\neq j}} \left\{\mathcal{X}_{i',j'}^{c}(t), N_{i',j'}(t) \ge (\log\log T)^{1/3}\right\}, \mathcal{Y}_{i,j}(t), \mathcal{Z}_\delta^{c}(t)\right] \\ &+ \sum_{(i',j')\in\mathcal{P}_{i\neq j}} \sum_{t=1}^{T} \mathbb{P}[N_{i',j'}(t) \le (\log\log T)^{1/3}, N_{i',j'}(t) \ge \alpha\sqrt{\log t}]. \qquad (30)\end{aligned}$$

Here,

$$\sum_{(i',j')\in\mathcal{P}_{i\neq j}} \sum_{t=1}^{T} \mathbb{P}[N_{i',j'}(t) \le (\log\log T)^{1/3}, N_{i',j'}(t) \ge \alpha\sqrt{\log t}] \le K^2 e^{\alpha^{-2}(\log\log T)^{2/3}} = o(\log T). \qquad (31)$$

Moreover,

$$\begin{aligned}&\sum_{t=1}^{T} \mathbb{P}\left[\bigcap_{(i',j')\in\mathcal{P}_{i\neq j}} \left\{\mathcal{X}_{i',j'}^{c}(t), N_{i',j'}(t) \ge (\log\log T)^{1/3}\right\}, \mathcal{Y}_{i,j}(t), \mathcal{Z}_\delta^{c}(t)\right] \\ &\le \sum_{n=1}^{T} \mathbb{P}\left[\bigcup_{t=n}^{T} \left\{|\hat{\mu}_{i,j}(t) - 1/2| \ge \beta/\log\log t, \mathcal{Y}_{i,j}(t), \mathcal{Z}_\delta^{c}(t), \bigcap_{(i',j')\in\mathcal{P}_{i\neq j}} N_{i',j'}(t) \ge (\log\log T)^{1/3}, N_{i,j}(t) = n\right\}\right] \\ &\le \sum_{n=1}^{\log T((\log\log T/\beta)^2/2)} \mathbb{P}\left[\bigcup_{t=1}^{T} \left\{\mathcal{Z}_\delta^{c}(t), \bigcap_{(i',j')\in\mathcal{P}_{i\neq j}} N_{i',j'}(t) \ge (\log\log T)^{1/3}, N_{i,j}(t) = n\right\}\right] \\ &\qquad \text{(by } \mathcal{R}_{i_1}(\{\nu_{i,j}\}), \mathcal{R}_{i_1}^{\mathrm{E}}(\{\nu_{i,j}\}) \subset [0, 1/d_{\mathrm{KL}}(\nu_{i,j}, 1/2)]^{K(K-1)/2} \text{ and Pinsker's inequality)} \\ &\le e^{-\Omega((\log\log T)^{1/3})} O((\log T)(\log\log T)^2) = o(\log T), \qquad (32)\end{aligned}$$

where we used the fact that

$$\begin{aligned}\mathbb{P}\left[\bigcup_{t=1}^{T} \left\{\mathcal{Z}_\delta^{c}(t), \bigcap_{(i',j')\in\mathcal{P}_{i\neq j}} N_{i',j'}(t) \ge (\log\log T)^{1/3}\right\}\right] &\le \sum_{(i',j')\in\mathcal{P}_{i\neq j}} \sum_{n=(\log\log T)^{1/3}}^{T} \mathbb{P}[|\hat{\mu}_{i,j}^{n} - \mu_{i,j}| > \delta] \\ &\le \sum_{(i',j')\in\mathcal{P}_{i\neq j}} \sum_{n=(\log\log T)^{1/3}}^{T} 2\mathrm{e}^{-2n\delta} = \mathrm{e}^{-\Omega((\log\log T)^{1/3})}.\end{aligned}$$

Combining (30), (31), and (32) completes the proof. □