Position-based Multiple-play Bandit Problem with Unknown Position Bias

Junpei Komiyama
The University of Tokyo
junpei@komiyama.info

Junya Honda
The University of Tokyo / RIKEN
honda@stat.t.u-tokyo.ac.jp

Akiko Takeda
The Institute of Statistical Mathematics / RIKEN
atakeda@ism.ac.jp

Abstract

Motivated by online advertising, we study a multiple-play multi-armed bandit problem with position bias that involves several slots and the latter slots yield fewer rewards. We characterize the hardness of the problem by deriving an asymptotic regret bound. We propose the Permutation Minimum Empirical Divergence (PMED) algorithm and derive its asymptotically optimal regret bound. Because of the uncertainty of the position bias, the optimal algorithm for such a problem requires non-convex optimizations that are different from usual partial monitoring and semi-bandit problems. We propose a cutting-plane method and related bi-convex relaxation for these optimizations by using auxiliary variables.

1 Introduction

One of the most important industries related to computer science is online advertising. In the United States, 72.5 billion dollars was spent on online advertising [19] in 2016. Most online advertising is viewed on web pages during Internet browsing. A web-site owner has a set of possible advertisements (ads): some of them are more attractive than others, and the owner would like to maximize the attention of visiting users. One of the observable metrics of the user attention is the number of clicks on the ads. By considering each ad (resp. click) to be an arm (resp. reward) and assuming only one slot is available for advertisements, the maximization of clicks boils down to the so-called multi-armed bandit problem, where the arm with the largest expected reward is sought.

When two or more ad slots are available on the web page, the problem boils down to a multiple-play multi-armed bandit problem. Several variants of the multiple play bandit problem and its extension called semi-bandit problem have been considered in the literature. Arguably, the simplest is one assuming that an ad receives equal clicks regardless of its position [2, 24]. In practice, ads receive less clicks when they are placed at bottom slots; this is so-called position bias.

A well-known model that explains position bias is the cascade model [23], which assumes that the users' attention goes from top to bottom until they lose interest. While this model explains position bias in early positions well [10], a drawback to the cascade model when it is applied to the bandit setting [26] is that the order of the allocated ads does not affect the reward, which is not very natural. To resolve this issue, Combes et al. [8] introduced a weight for each slot that corresponds to the reward obtained by clicking on that slot. However, no principled way of defining the weight has been described.

An extension of the cascade model, called the dependent click model (DCM) [14], addresses these issues by admitting multiple clicks of a user. In DCM, each slot is associated with a probability that

the user loses interest in the following ads if the current ad is interesting. While the algorithm in Katariya et al. [21] cleverly exploits this structure, it still depends on the cascade assumption, and as a result it discards some of the feedback on the latter slots, which reduces the efficiency of the algorithm. Moreover, the reward in DCM does not exactly correspond to the number of clicks.

Lagrée et al. [27] has studied a position-based model (PBM) where each slot has its own discount factor on the number of clicks. PBM takes the order of the shown ads into consideration. However, the algorithms proposed in Lagrée et al. [27] are "half-online" in the sense that the value of an ad is adaptively estimated, whereas the values of the slots are estimated by using an off-line dataset. Such an off-line computation is not very handy since the click trend varies depending on the day and hour [1]. Moreover, a significant portion of online advertisements is sold via ad networks [34]. As a result, advertisers have to deal with thousands of web pages to show their ads. Taking these aspects into consideration, pre-computing position bias for each web page limits the use of these algorithms.

To address this issue, we provide a way to allocate advertisements in a fully online manner by considering "PBM under Uncertainty of position bias" (PBMU). One of the challenges when the uncertainty of a position-based factor is taken into account is that, when some ad appears to have a small click through rate (CTR, the probability of click) in some slot, we cannot directly attribute it to either the arm or the slot. In this sense, several combinations of ads and slots need to be examined to estimate both the ad-based and position-based model parameters.

Note also that an extension of the non-stochastic bandit approach [3] to multiple-play, such as the ordered slate model [20], is general enough to deal with PBMU. However, algorithms based on the non-stochastic approach do not always perform well in compensation for its generality. Another extension of multi-armed bandit problems is the partial monitoring problem [31, 4] that admits the case in which the parameters are not directly observable. However, partial monitoring is inefficient at solving bandit problems: a K-armed bandit problem with binary rewards corresponds to a partial monitoring problem with 2^K possible outcomes. As a result, the existing partial monitoring algorithms, such as the ones in [33, 25], are not practical even for a moderate number of arms. Besides, the computation of a feasible solution in PBMU requires non-convex optimizations as we will see in Section 5. This implies that PBMU cannot directly be converted into the partial monitoring where such a non-convex optimization does not appear [25].

The contributions of this paper are as follows: First, we study the position-based bandit model with uncertainty (PBMU) and derive a regret lower bound (Section 3). Second, we propose an algorithm that efficiently utilizes feedback (Section 4). One of the challenges in the multiple-play bandit problem is that there is an exponentially large number of possible sequences of arms to allocate at each round. We reduce the number of candidates by using a bipartite matching algorithm that runs in a polynomial time to the number of arms. The performance of the proposed algorithm is verified in Section 6. Third, a slightly modified version of the algorithm is analyzed in Section 7. This algorithm has a regret upper bound that matches the lower bound. Finally, we reveal that the lower bound is related to a linear optimization problem with an infinite number of constraints. Such an optimization problem appears in many versions of the bandit problem [9, 25, 12]. We propose an optimization method that reduces it to a finite-constraint linear optimization based on a version of the cutting-plane method (Section 5). Related non-convex optimizations that are characteristic to PBMU are solved by using bi-convex relaxation. Such optimization methods are of interest in solving even larger classes of bandit problems.

2 Problem Setup

Let K be the number of arms (ads) and $L < K$ be the number of slots. Each arm $i \in [K] = \{1, 2, \ldots, K\}$ is associated with a distinct parameter $\theta_i^* \in (0, 1)$, and each slot $l \in [L]$ is associated with a parameter $\kappa_l^* \in (0, 1]$. At each round $t = 1, 2, \ldots, T$, the system selects L arms $I(t) = (I_1(t), \ldots, I_L(t))$ and receives a corresponding binary reward (click or non-click) for each slot. The reward of the l-th slot is i.i.d. drawn from a Bernoulli distribution $\mathrm{Ber}(\mu_{I_l(t),l}^*)$, where $\mu_{i,l}^* = \theta_i^* \kappa_l^*$. Although the slot-based parameters are unknown, it is natural that the ads receives more clicks when they are placed at early slots: we assume $\kappa_1^* > \kappa_2^* > \cdots > \kappa_L^* > 0$ and this order is known.

Note that this model is redundant: a model with $\mu_{i,l}^* = \theta_i^* \kappa_l^*$ is equivalent to the model with $\mu_{i,l}^* = (\theta_i^*/\kappa_1)(\kappa_l^* \kappa_1)$. Therefore, without loss of generality, we assume $\kappa_1 = 1$. In summary,

2

this model involves $K + L$ parameters $\{\theta_i^*\}_{i \in [K]}$ and $\{\kappa_l^*\}_{l \in [L]}$, and the number of rounds T. The parameters except for $\kappa_1 = 1$ are unknown to the system. Let $N_{i,l}(t)$ be the number of rounds before t-th round at which arm i was in slot l (i.e., $N_{i,l}(t) = \sum_{t'=1}^{t-1} \mathbf{1}\{i = I_l(t')\}$, where $\mathbf{1}\{\mathcal{E}\}$ is 1 if $\mathcal{E}$ holds and 0 otherwise). In the following, we abbreviate arm i in slot l to "pair (i, l)". Let $\hat{\mu}_{i,l}(t)$ be the empirical mean of the reward of pair (i, l) after the first $t - 1$ rounds.

The goal of the system is to maximize the cumulative rewards by using some sophisticated algorithm. Without loss of generality, we can assume $\theta_1^* > \theta_2^* > \theta_3^* > \cdots > \theta_K^*$. The algorithm cannot exploit this ordering. In this model, allocating arms of larger expected rewards on earlier slots increases expected rewards: As a result, allocating arms $1, 2, \ldots, L$ to slots $1, 2, \ldots, L$ maximizes the expected reward. A quantity called (pseudo) regret is defined as: $\mathrm{Reg}(T) = \sum_{t=1}^T \left(\sum_{i \in [L]} (\theta_i^* - \theta_{I_i(t)}^*) \kappa_i^* \right)$, and $\mathbb{E}[\mathrm{Reg}(T)]$ is used for evaluating the performance of an algorithm. Let $\Delta_{i,l} = \theta_l^* \kappa_l^* - \theta_i^* \kappa_l^*$. Regret can be alternatively represented as $\mathrm{Reg}(T) = \sum_{(i,l) \in [K] \times [L]} \Delta_{i,l} N_{i,l}(T)$. The regret increases unless $I(t) = (1, 2, \ldots, L)$.

3 Regret Lower Bound

Here, we derive an asymptotic regret lower bound when $T \to \infty$. In the context of the standard multi-armed bandit problem, Lai and Robbins [28] derived a regret lower bound for strongly consistent algorithms, and it is followed by many extensions, such as the one for multi-parameter distributions [6] and the ones for Markov decision processes [13, 7]. Intuitively, a strongly consistent algorithm is "uniformly good" in the sense that it works well with any set of model parameters. Their result was extended to the multiple-play [2] and PBM [27] cases. We further extend it to the case of PBMU.

Let $\mathcal{T}_{\mathrm{all}} = \{(\theta_1', \ldots, \theta_K') \in (0,1)^K\}$ and $\mathcal{K}_{\mathrm{all}} = \{(\kappa_1', \ldots, \kappa_L') : 1 = \kappa_1' > \kappa_2' > \cdots > \kappa_L' > 0\}$ be the sets of all possible values on the parameters of the arms and slots, respectively. Let $(1), \ldots, (K)$ be a permutation of $1, \ldots, K$ and $\mathcal{T}_{(1),\ldots,(L)}$ be the subset of $\mathcal{T}_{\mathrm{all}}$ such that the i-th best arm is (i). Namely,

$$\mathcal{T}_{(1),\ldots,(L)} = \left\{ (\theta_1', \ldots, \theta_K') \in (0,1)^K : \theta_{(1)}' > \theta_{(2)}' > \cdots > \theta_{(L)}', \forall_{i \notin \{(1),\ldots,(L)\}} (\theta_i' < \theta_{(L)}') \right\},$$

and $\mathcal{T}_{(1),\ldots,(L)}^c = \mathcal{T}_{\mathrm{all}} \setminus \mathcal{T}_{(1),\ldots,(L)}$. An algorithm is *strongly consistent* if $\mathbb{E}[\mathrm{Reg}(T)] = o(T^a)$ for any $a > 0$ given any instance of the bandit problem with its parameters $\{\theta_i'\}_{i \in [K]} \in \mathcal{T}_{\mathrm{all}}, \{\kappa_l'\} \in \mathcal{K}_{\mathrm{all}}$. The following lemma, whose proof is in Appendix F, lower-bounds the number of draws on the pairs of arms and slots.

Lemma 1. (Lower bound on the number of draws) *The following inequality holds for $N_{i,l}(T)$ of the strongly consistent algorithm:*

$$\forall_{\{\theta_i'\} \in \mathcal{T}_{1,\ldots,L}^c, \{\kappa_l'\} \in \mathcal{K}_{\mathrm{all}}} \sum_{(i,l) \in [K] \times [L]} \mathbb{E}[N_{i,l}(T)] d_{\mathrm{KL}}(\theta_i^* \kappa_l^*, \theta_i' \kappa_l') \geq \log T - o(\log T),$$

where $d_{\mathrm{KL}}(p, q) = p \log(p/q) + (1 - p) \log((1 - p)/(1 - q))$ is the KL divergence between two Bernoulli distributions.

Such a divergence-based bound appears in many stochastic bandit problems. However, unlike other bandit problems, the argument inside the KL divergence is a product of parameters $\theta_i' \kappa_l'$: While $d_{\mathrm{KL}}(\cdot, \theta_i' \kappa_l')$ is convex to $\theta_i' \kappa_l'$, it is not convex to the parameter space $\{\theta_i'\}, \{\kappa_l'\}$. Therefore, finding a set of parameters that minimizes $\sum_{i,l} d_{\mathrm{KL}}(\mu_{i,l}, \theta_i' \kappa_l')$ is non-convex, which makes PBMU difficult.

Furthermore, we can formalize the regret lower bound in what follows. Let

$$\mathcal{Q} = \left\{ \{q_{i,l}\} \in [0, \infty)^{[K] \times [K]} : \forall_{i \in [K-1]} \sum_{l \in [K]} q_{i,l} = \sum_{l \in [K]} q_{i+1,l}, \forall_{l \in [K-1]} \sum_{i \in [K]} q_{i,l} = \sum_{i \in [K]} q_{i,l+1} \right\}.$$

Intuitively, $\{q_{i,l}\}$ for $l \leq L$ corresponds to the draw of arm i in slot l, and $\{q_{i,l}\}$ for $l > L$ corresponds to the non-draw of arm i, as we will see later. The following quantities characterizes the minimum

3

amount of exploration for consistency:

$$\mathcal{R}_{(1),\ldots,(L)}(\{\mu_{i,l}\}, \{\theta_i\}, \{\kappa_l\}) = \left\{ \{q_{i,l}\} \in \mathcal{Q} : \inf_{\{\theta_i'\} \in \mathcal{T}_{(1),\ldots,(L)}^c, \{\kappa_l'\} \in \mathcal{K}_{\mathrm{all}} : \forall_{i \in [L]} \theta_i' \kappa_i' = \theta_i \kappa_i} \right.$$
$$\left. \sum_{(i,l) \in [K] \times [L] : i \neq (l)} q_{i,l} d_{\mathrm{KL}}\left(\mu_{i,l}, \theta_i' \kappa_l'\right) \geq 1 \right\}. \quad (1)$$

Equality (1) states that drawing each pair (i, l) for $N_{i,l} = q_{i,l} \log T$ times suffices to reduce the risk that the true parameter is $\{\theta_i'\}, \{\kappa_l'\}$ for any parameters $\{\theta_i'\}, \{\kappa_l'\}$ such that $\theta_i' \in \mathcal{T}_{(1),\ldots,(L)}^c$ and $\theta_i' \kappa_l' = \theta_i \kappa_i$ for any $i \in [L]$. Note that the constraint $\theta_i' \kappa_l' = \theta_i \kappa_i$ corresponds to the fact that drawing an optimal list of arms does not increase the regret: Intuitively, this corresponds to the fact that the true parameter of the best arm is obtained for free in the regret lower bound of the standard bandit problem[1]. Moreover, let

$$C_{(1),\ldots,(L)}^*(\{\mu_{i,l}\}, \{\theta_i\}, \{\kappa_l\}) = \inf_{\{q_{i,l}\} \in \mathcal{R}_{(1),\ldots,(L)}(\{\mu_{i,l}\}, \{\theta_i\}, \{\kappa_l\})} \sum_{(i,l) \in [K] \times [L]} \Delta_{i,l} q_{i,l},$$

the set of optimal solutions of which is denoted by

$$\mathcal{R}_{(1),\ldots,(L)}^*(\{\mu_{i,l}\}, \{\theta_i\}, \{\kappa_l\}) = \left\{ \{q_{i,l}\} \in \mathcal{R}_{(1),\ldots,(L)}(\{\mu_{i,l}\}, \{\theta_i\}, \{\kappa_l\}) : \right.$$
$$\left. \sum_{(i,l) \in [K] \times [L]} \Delta_{i,l} q_{i,l} = C_{(1),\ldots,(L)}^*(\{\mu_{i,l}\}, \{\theta_i\}, \{\kappa_l\}) \right\}. \quad (2)$$

The value $C_{1,\ldots,L}^* \log T$ is the possible minimum regret such that the minimum divergence of $\{\theta_i^*\}, \{\kappa_l^*\}$ from any $\{\theta_i'\}, \{\kappa_l'\}$ is larger than $\log T$. Using Lemma 1 yields the following regret lower bound, whose proof is also in the Appendix F.

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

$$\mathbb{E}[\mathrm{Reg}(T)] \geq C_{1,\ldots,L}^*(\{\mu_{i,l}^*\}, \{\theta_i^*\}, \{\kappa_l^*\}) \log T - o(\log T).$$

Remark 3. $N_{i,l} = (\log T)/d_{\mathrm{KL}}(\theta_i^* \kappa_i^*, \theta_j^* \kappa_i^*)$ for $j = \min(i - 1, L)$ satisfies the conditions in Lemma 1, which means that regret lower bound in Theorem 2 is $O(K \log T / \Delta) = O(K \log T)$, where $\Delta = \min_{i \neq j, l \neq m} |\theta_i^* - \theta_j^*||\kappa_l^* - \kappa_m^*|$.

4 Algorithm

Our algorithm, called Permutation Minimum Empirical Divergence (PMED), is closely related to the optimization we discussed in Section 3.

4.1 PMED Algorithm

We denote a list of L arms that are drawn at each round as L-allocation. For example, $(3, 2, 1, 5)$ is a 4-allocation, which corresponds to allocating arms $3, 2, 1, 5$ to slots $1, 2, 3, 4$, respectively. Like the Deterministic Minimum Empirical Divergence (DMED) algorithm [17] for the single-play multi-armed bandit problem, Algorithm 1 selects arms by using a loop. $L_C = L_C(t)$ is the set of L-allocations in the current loop, and $L_N = L_N(t)$ is the set of L-allocations that are to be drawn in the next loop. Note that, $|L_N| \geq 1$ always holds at the end of each loop so that at least one element is

---

[1]The infimum should take parameters $\theta_i' \kappa_i' \neq \theta_i \kappa_i$ into consideration. However, such parameters can be removed without increasing regret, and thus the infimum over $\theta_i' \kappa_i' = \theta_i \kappa_i$ suffices. This can be understood because the regret bound of the standard K-armed bandit problem with expectation of each arm μ_i is $\sum_{i=2}^{K} (\log T)/d_{\mathrm{KL}}(\mu_i, \mu_1)$: Arm 1 is drawn without increasing regret, and thus estimation of μ_1 can be arbitrary accurate. In our case placing arms $1, \ldots, L$ into slots $1, \ldots, L$ does not increase the regret, and thus the estimation of the product parameter $\theta_i \kappa_i$ for each $i \in [L]$ is very accurate.

Algorithm 1 PMED and PMED-Hinge Algorithms

1: Input: $\alpha > 0$, $\beta > 0$ (for PMED-Hinge), $f(n) = \gamma/\sqrt{n}$ with $\gamma > 0$ (for PMED-Hinge).
2: $L_N \leftarrow \emptyset$. $L_C \leftarrow \{v_1^{\mathrm{mod}}, \ldots, v_K^{\mathrm{mod}}\}$.
3: **while** $t \le T$ **do**
4:     **for** each $v_m^{\mathrm{mod}} : m \in [K]$ **do**
5:         If there exists some pair $(i, l) \in v_m^{\mathrm{mod}}$ such that $N_{i,l}(t) < \alpha\sqrt{\log t}$, then put v_m^{mod} into L_N.
6:     **end for**
7:     Compute the MLE $\{\hat{\theta}_i(t)\}_{i=1}^K$, $\{\hat{\kappa}_l(t)\}_{l=1}^L$

$$= \begin{cases} \min_{\{\theta_i, \kappa_l\}} \sum_{(i,l) \in [K] \times [L]} N_{i,l}(t) d_{\mathrm{KL}}(\hat{\mu}_{i,l}(t), \theta_i \kappa_l) & \text{(PMED)} \\ \min_{\{\theta_i, \kappa_l\}} \sum_{(i,l) \in [K] \times [L]} N_{i,l}(t) \left(d_{\mathrm{KL}}(\hat{\mu}_{i,l}(t), \theta_i \kappa_l) - f(N_{i,l}(t))\right)_+. & \text{(PMED-Hinge)} \end{cases}$$

8:     **if** Algorithm is PMED-Hinge **then**
9:         If $|\hat{\theta}_i(t) - \hat{\theta}_j(t)| < \beta/(\log \log t)$ for some $i \ne j$ or $|\hat{\kappa}_l(t) - \hat{\kappa}_m(t)| < \beta/(\log \log t)$ for some $l \ne m$, then put all of $v_1^{\mathrm{mod}}, \ldots, v_K^{\mathrm{mod}}$ to L_N.
10:         If $\bigcup_{(i,l) \in [K] \times [L]} \{d_{\mathrm{KL}}(\hat{\mu}_{i,l}(t), \hat{\theta}_i(t)\hat{\kappa}_l(t)) > f(N_{i,l}(t))\}$ holds, then put all of $v_1^{\mathrm{mod}}, \ldots, v_K^{\mathrm{mod}}$ into L_N.
11:     **end if**
12:     Compute $\{q_{i,l}\} \in \begin{cases} \mathcal{R}^*_{\hat{1}(t), \ldots, \hat{L}(t)}(\{\hat{\mu}_{i,l}(t)\}, \{\hat{\theta}_i(t)\}, \{\hat{\kappa}_l(t)\}) & \text{(PMED)} \\ \mathcal{R}^{*,\mathrm{H}}_{\hat{1}(t), \ldots, \hat{L}(t)}(\{\hat{\mu}_{i,l}(t)\}, \{\hat{\theta}_i(t)\}, \{\hat{\kappa}_l(t)\}, \{f(N_{i,l}(t))\}). & \text{(PMED-Hinge)} \end{cases}$
13:     $\tilde{N}_{i,l} \leftarrow q_{i,l} \log t$ for each $(i, l) \in [K] \times [K]$.
14:     Decompose $\tilde{N}_{i,l} = \sum_v c_v^{\mathrm{req}} e_v$ where e_v for each v is a permutation matrix and $c_v^{\mathrm{req}} > 0$ by using Algorithm 2.
15:     $r_{i,l} \leftarrow N_{i,l}(t)$.
16:     **for** each permutation matrix e_v **do**
17:         $c_v^{\mathrm{aff}} \leftarrow \min\left(c_v^{\mathrm{req}}, \max_c \left\{c > 0 : \min_{(i,l) \in [K] \times [L]}(r_{i,l} - c\, e_{v,i,l}) \ge 0\right\}\right)$.
18:         Let $(v_1, \ldots, v_L)$ be the L-allocation corresponding to e_v. If $c_v^{\mathrm{aff}} < c_v^{\mathrm{req}}$ and there exists a pair (v_l, l) that is in none of the L-allocations in L_N, then put $(v_1, \ldots, v_L)$ into L_N.
19:         $r_{i,l} \leftarrow r_{i,l} - c_v^{\mathrm{aff}} e_{v,i,l}$.
20:     **end for**
21:     Select $I(t) \in L_C$ in an arbitrary fixed order. $L_C \to L_C \setminus \{I(t)\}$.
22:     Put $(\hat{1}(t), \ldots, \hat{L}(t))$ into L_N.
23:     If $L_C = \emptyset$ then $L_C \leftarrow L_N$, $L_N \leftarrow \emptyset$.
24: **end while**
---

put into L_C. There are three lines where L-allocations are put into L_N without duplication: Lines 5, 18, and 22. We explain each of these lines below.

Line 5 is a uniform exploration over all pairs (i, l). For $m \in [K]$, let v_m^{mod} be an L-allocation $(1 + \mathrm{mod}_K(m), 1 + \mathrm{mod}_K(1 + m), \ldots, 1 + \mathrm{mod}_K(L + m - 1))$, where $\mathrm{mod}_K(x)$ is the minimum non-negative integer among $\{x - cK : c \in \mathbb{N}\}$. From the definition of v_m^{mod}, any pair $(i, l) \in [K] \times [L]$ belongs to exactly one of $v_1^{\mathrm{mod}}, \ldots, v_K^{\mathrm{mod}}$. If some pair (i, l) is not allocated $\alpha\sqrt{\log t}$ times, a corresponding L-allocation is put into L_N. This exploration stabilizes the estimators.

Line 18 and related routines are based on the optimal amount of explorations. $\{\tilde{N}_{i,l}\}_{i \in [K], l \in [K]}$ is calculated by plugging in the maximum likelihood estimator (MLE) $(\{\hat{\theta}_i\}_{i \in [K]}, \{\hat{\kappa}_l\}_{l \in [L]})$ into the optimization problem of Inequality (2). As $\{\tilde{N}_{i,l}\}$ is a set of $K \times K$ variables[2], the algorithm needs to convert it into a set of L-allocations to put them into L_N. This is done by decomposing it into a set of permutation matrices, which we will explain in Section 4.2.

Line 22 is for exploitation: If no pair is put to L_N by Line 5 or Line 18 and L_C is empty, then Line 22 puts arms $(\hat{1}(t), \ldots, \hat{L}(t))$ of the top-L largest $\{\hat{\theta}_i(t)\}$ (with ties broken arbitrarily) into L_N.

---

[2]$K \times K$ is not a typo of $K \times L$: $\{q_{i,l}\}$ and $\{\tilde{N}_{i,l}\}$ are sets of K^2 variables.

Algorithm 2 Permutation Matrix Decomposition

1: Input: $N_{i,l}$.
2: $\bar{N}_{i,l} \leftarrow N_{i,l}$.
3: **while** $\bar{N}_{i,l} > 0$ for some $(i,l) \in [K] \times [K]$ **do**
4:     Find a permutation matrix e_v such that, for any i, l such that $e_{v,i,l} = 1 \Rightarrow \bar{N}_{i,l} > 0$.
5:     Let $c_v^{\text{req}} = \max_c \left\{ c > 0 : \min_{(i,l) \in [K] \times [K]} (\bar{N}_{i,l} - c e_{v,i,l}) \geq 0 \right\}$.
6:     $\bar{N}_{i,l} \leftarrow \bar{N}_{i,l} - c_v^{\text{req}} e_{v,i,l}$ for each $(i,l) \in [K] \times [K]$.
7: **end while**
8: Output $\{c_v^{\text{req}}, e_v\}$
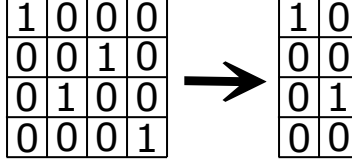
Figure 1: A permutation matrix with $K = 4$, where $(i,l) = 1$ for $(i,l) \in (1,1), (2,3), (3,2), (4,4)$. If $L = 2$, this matrix corresponds to allocating arm 1 in slot 1 and arm 3 in slot 2.

4.2 Permutation Matrix and Allocation Strategy

In this section, we discuss the way to convert $\{\tilde{N}_{i,l}\} = \{q_{i,l} \log t\}$, the estimated optimal amount of exploration, into L-allocations. A permutation matrix is a square matrix that has exactly one entry of 1 at each row and each column and 0s elsewhere (Figure 1, left). There are $K!$ permutation matrices since they corresponds to ordering K elements. Therefore, even though $\{q_{i,l}\}$ can be obviously decomposed into a linear combination of permutation matrices, it is not clear how to compute them without computing the set of all permutation matrices that are exponentially large in K. Algorithm 2 solves this problem: Let $\bar{N}_{i,l}$ be a temporal variable that is initialized by $\tilde{N}_{i,l}$ at the beginning. In each iteration, it subtracts a scalar multiplication of a permutation matrix e_v whose (i,l) entry $e_{v,i,l}$ of value 1 corresponds to $\bar{N}_{i,l} > 0$. (Line 6 in Algorithm 2). This boils down to finding a perfect matching in a bipartite graph where the left (resp. right) nodes correspond to rows (resp. columns) and edges between nodes i and l are spanned if $\bar{N}_{i,l} > 0$. Although a naive greedy fails in such a matching problem (c.f., Appendix A), a maximal matching in a bipartite graph can be computed by the Hopcroft–Karp algorithm [18] in $O(K^{2.5})$ times, and Theorem 4 below ensures that the maximum matching is always perfect:

Theorem 4. (Existence of a perfect matching) *For any* $\{\bar{N}_{i,l} \in [K] \times [K] : \bar{N}_{i,l} \geq 0, \exists_{(i,l)} \bar{N}_{i,l} > 0\}$ *such that the sums of each row and column are equal, there exists a permutation matrix* e_v *such that* $\forall_{(i,l) \in [K] \times [K]: e_{v,i,l} = 1} \bar{N}_{i,l} > 0$.

The proof of Theorem 4 is in Appendix E. Each subtraction increases the number of 0 entries in $\bar{N}_{i,l}$ (Line 5 in Algorithm 2); Algorithm 2 runs in $O(K^{4.5})$ times by computing at most $O(K^2)$ perfect matching sub-problems, and as a result it decomposes $\tilde{N}_{i,l}$ into a positive linear combination of permutation matrices. The main algorithm checks whether each the entries of the permutation matrices are sufficiently explored (Line 18 in Algorithm 1), and draws an L-allocation corresponding to a permutation matrix (Figure 1, right) if under-explored.

5 Optimizations

This section discusses two optimizations that appear in Algorithm 1, namely, the MLE computation (Line 7), and the computation of the optimal solution (Line 12).

MLE (Line 7) is the solution of a bi-convex optimization: the optimization of $\{\theta_i\}$ (resp. $\{\kappa_l\}$) is convex when we view $\{\kappa_l\}$ (resp. $\{\theta_i\}$) as a constant. Therefore, off-the-shelf tools for optimizing convex functions (e.g., Newton's method) are applicable to alternately optimizing $\{\theta_i\}$ and $\{\kappa_l\}$. Assuming that each convex optimization yields an optimal value, such an alternate optimization

---

Algorithm 3 Cutting-plane method for obtaining $\{q_{i,l}\}$ on Line 12 of Algorithm 1

1: Input: the number of iterations S, nominal constraint $\{\theta_i^{(0)}\} \in \mathcal{T}_{\hat{1}(t),\ldots,\hat{L}(t)}^c$.

2: **for** $s = 1, 2, \ldots, S$ **do**

3:    Find $q_{i,l}^{(s)} \leftarrow \min_{\{q_{i,l}\} \in \mathcal{Q}} \sum_{(i,l) \in [K] \times [L]} \Delta_{i,l} q_{i,l}$ such that

$$\sum_{(i,l) \in [K] \times [L]: i \neq \hat{l}(t)} q_{i,l} d_{\mathrm{KL}} \left(\hat{\mu}_{i,l}(t), \theta_i' \frac{\hat{\theta}_l(t) \hat{\kappa}_l(t)}{\theta_l'} \right) \geq 1$$

     for all $\{\theta_i'\} \in \{\theta_i^{(0)}\}, \{\theta_i^{(1)}\}, \ldots, \{\theta_i^{(s-1)}\}$.

4:    Find $\{\theta_i^{(s)}\} \leftarrow \min_{\{\theta_i'\}} \sum_{(i,l) \in [K] \times [L]} q_{i,l}^{(s)} d_{\mathrm{KL}}(\hat{\mu}_{i,l}(t), \theta_i' \frac{\hat{\theta}_l(t) \hat{\kappa}_l(t)}{\theta_l'})$.

5: **end for**

---

monotonically decreases the objective function and thus converges. Note that a local minimum obtained by bi-convex optimizations is not always a global minimum due to its non-convex nature.

Although the computation of the optimal solution (Line 12) involves $\{\theta_i'\}$ and $\{\kappa_l'\}$, the constraint eliminates latter variables as $\kappa_i' = \hat{\theta}_i(t)\hat{\kappa}_i(t)/\theta_i'$. This optimization is a linear semi-infinite programming (LSIP) on $\{q_{i,l}\}$, which is a linear programming (LP) with an infinite set of linear constraints parameterized by $\{\theta_i'\}$. Algorithm 3 is the cutting-plane method with pessimistic oracle [29] that boils the LSIP down to finite constraint LPs. At each iteration s, it adds a new constraint $\{\theta_i^{(s)}\} \in \mathcal{T}_{\hat{1}(t),\ldots,\hat{L}(t)}^c$ that is "hardest" in a sense that it minimizes the sum of divergences (Line 4 in Algorithm 3). The following theorem guarantees the convergence of the algorithm when the exactly hardest constraint is found.

Theorem 5. (Convergence of the cutting-plane method, Mutapcic and Boyd [29, Section 5.2]) *Assume that there exists a constant C and that the constraint $f(\{\theta_i'\}) = \sum_{(i,l) \in [K] \times [L]} q_{i,l}^{(s)} d_{\mathrm{KL}}(\hat{\mu}_{i,l}(t), \theta_i' \frac{\hat{\theta}_l(t)\hat{\kappa}_l(t)}{\theta_l'})$ is Lipchitz continuous as $|f(\{\theta_i^{(1)}\}) - f(\{\theta_i^{(2)}\})| \leq C||\{\theta_i^{(1)}\} - \{\theta_i^{(2)}\}||$, where the norm $||\cdot||$ is any L_p norm. Then, Algorithm 3 converges to its optimal solution as $S \to \infty$.*

Although the Lipchitz continuity assumption does not hold as $d_{\mathrm{KL}}(p, q)$ approaches infinity when q is close to 0 or 1, by restricting q to some region $[\epsilon, 1 - \epsilon]$, Lipchitz continuity can be guaranteed for some $C = C(\epsilon)$. Theorem 5 assumes the availability of an exact solution to the hardest constraint, which is generally hard since this objective is non-convex in its nature. Still, we can obtain a fair solution with the following reasons: First, although the space $\mathcal{T}_{\hat{1}(t),\ldots,\hat{L}(t)}^c$ is not convex, it suffices to consider each of the convex subspaces $\left\{ \{\theta_i'\} \in (0,1)^K : \theta_{\hat{1}(t)}' \geq \cdots \geq \theta_{\hat{L}(t)}', \theta_{\hat{X}(t)}' = \theta_{\hat{l}(t)}' \right\}$ where $X = \min(L, l - 1)$, for each $l \in [K] \setminus \{1\}$ separately because the hardest constraint is always in one of these subspaces (which follows from the convexity of the objective function). Second, the following bi-convex relaxation can be used: Let $\eta_1', \ldots, \eta_L'$ be auxiliary variables that correspond to $1/\theta_1', \ldots, 1/\theta_L'$. Namely, we optimize a relaxed objective function $\sum_{(i,l) \in [K] \times [L]} \left(q_{i,l}^{(s)} d_{\mathrm{KL}}(\hat{\mu}_{i,l}(t), \theta_i' \eta_l' \hat{\theta}_l(t) \hat{\kappa}_l(t)) \right) + \phi \sum_{i \in [L]} (\theta_i' \eta_i' - 1)^2$, where $\phi > 0$ is a penalty parameter. Convexity of KL divergence implies that this objective is a bi-convex function of $\{\theta_i'\}$ and $\{\eta_l'\}$, and thus an alternate optimization is effective. Setting $\phi \to \infty$ induces a solution in which η_i' is equal to $1/\theta_i'$ ([30, Theorem 17.1]). Our algorithm starts with a small value of ϕ; then it gradually increases ϕ.

6 Experiment

To evaluate the empirical performance of the proposed algorithms, we conducted computer simulations with synthetic and real-world datasets. The compared algorithms are MP-TS [24], dcmKL-UCB [21], PBM-PIE [27], and PMED (proposed in this paper). MP-TS is an algorithm based on Thompson sampling [32] that ignores position bias: it draws the top-L arms on the basis of posterior sampling, and the posterior is calculated without considering position bias. DcmKL-UCB is a KL-UCB [11]
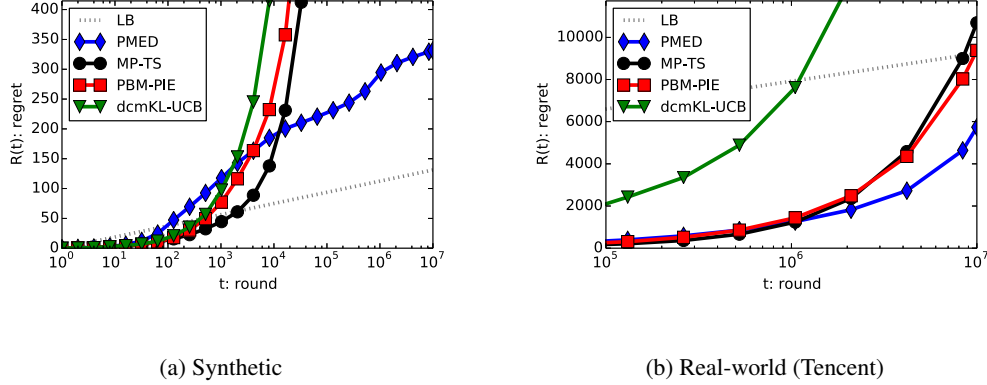
|(a) Synthetic | (b) Real-world (Tencent)|

Figure 2: Regret-round log-log plots of algorithms.

based algorithm that works under the DCM assumption. PBM-PIE is an algorithm that allocates top-$(L-1)$ slots greedily and allocates L-th arm based on the KL-UCB bound. Note that PBM-PIE requires an estimation of $\{\kappa_l^*\}$; here, a bi-convex optimization is used to estimate it[3]. We did not test PBM-TS [27], which is another algorithm for PBM, mainly because that its regret bound has not been derived yet. However, its regret appears to be asymptotically optimal when $\{\kappa_l^*\}$ are known (Figure 1(a) in Lagrée et al.[27]), and thus it does not explore sufficiently when there is uncertainty in the position bias. We set $\alpha = 10$ for PMED. We used the Gurobi LP solver[4] for solving the LPs. To speed up the computation, we skipped the bi-convex and LP optimizations in most rounds with large t and used the result of the last computation. We used the Newton's method (resp. a gradient method) for computing the MLE (resp. the hardest constraint) in Algorithm 3.

Synthetic data: This simulation was designed to check the consistency of the algorithms, and it involved 5 arms with $(\theta_1, \ldots, \theta_5) = (0.95, 0.8, 0.65, 0.5, 0.35)$, and 2 slots with $(\kappa_1, \kappa_2) = (1, 0.6)$. The experimental results are shown on the left of Figure 2. The results are averaged over 100 runs. LB is the simulated value of the regret lower bound in Section 3. While the regret of PMED converges, the other algorithms suffer a 100 times or larger regret than LB at $T = 10^7$, which implies that these algorithms are not consistent under our model.

Real-world data: Following the existing work [24, 27], we used the KDD Cup 2012 track 2 dataset [22] that involves session logs of soso.com, a search engine owned by Tencent. Each of the 150M lines from the log contains the user ID, the query, an ad, and a slot in $\{1, 2, 3\}$ at which the ad was displayed and a binary reward indicated (click/no-click). Following Lagrée et al. [27], we obtained major 8 queries. Using the click logs of the queries, the CTRs and position bias were estimated in order to maximize the likelihood by using bi-convex optimization in Section 4. Note that, the number of arms and parameters are slightly different from the ones reported previously [27]. For the sake of completeness, we show the parameters in Appendix C. We conducted 100 runs for each queries, and the right figure in Figure 2 shows the averaged regret over 8 queries. Although the gap between PMED and existing algorithms are not drastic compared with synthetic parameters, the existing algorithms suffer larger regret than PMED.

7 Analysis

Although the authors conjecture that PMED is optimal, it is hard to analyze it directly. The technically hardest part arises from the case in which the divergence of each action is small but not yet fully converged. To circumvent these difficulty, we devised a modified algorithm called PMED-Hinge (Algorithm 1) that involves extra exploration. In particular, we modify the optimization problem as

---

[3]The bi-convex optimization is identical to the one used for obtaining the MLE in PMED.
[4]http://www.gurobi.com

follows: Let

$$\mathcal{R}^{\mathrm{H}}_{(1),\dots,(L)}(\{\mu_{i,l}\},\{\theta_i\},\{\kappa_l\},\{\delta_{i,l}\}) = \Bigg\{ \{q_{i,l}\} \in \mathcal{Q} : \inf_{\{\theta'_i\}\in\mathcal{T}^{c}_{(1),\dots,(L)},\{\kappa'_l\}\in\mathcal{K}_{\mathrm{all}}:\forall_{l\in[L]}d_{\mathrm{KL}}(\mu_{(l),l},\theta'_{(l)}\kappa'_l)\leq\delta_{i,l}}$$

$$\sum_{(i,l)\in[K]\times[L]:i\neq(l)} q_{i,l}\left(d_{\mathrm{KL}}(\mu_{i,l},\theta'_i\kappa'_l) - \delta_{i,l}\right)_+ \geq 1 \Bigg\},$$

where $(x)_+ = \max(x,0)$. Moreover, let

$$C^{*,\mathrm{H}}_{(1),\dots,(L)}(\{\mu_{i,l}\},\{\theta_i\},\{\kappa_l\},\{\delta_{i,l}\}) = \inf_{\{q_{i,l}\}\in\mathcal{R}^{\mathrm{H}}_{(1),\dots,(L)}(\{\mu_{i,l}\},\{\theta_i\},\{\kappa_l\},\{\delta_{i,l}\})} \sum_{(i,l)\in[K]\times[L]} \Delta_{i,l}q_{i,l},$$

the optimal solution of which is

$$\mathcal{R}^{*,\mathrm{H}}_{(1),\dots,(L)}(\{\mu_{i,l}\},\{\theta_i\},\{\kappa_l\},\{\delta_{i,l}\}) = \Bigg\{ \{q_{i,l}\} \in \mathcal{R}^{\mathrm{H}}_{(1),\dots,(L)}(\{\mu_{i,l}\},\{\theta_i\},\{\kappa_l\},\{\delta_{i,l}\}) :$$

$$\sum_{(i,l)\in[K]\times[L]} \Delta_{i,l}q_{i,l} = C^{*,\mathrm{H}}_{(1),\dots,(L)}(\{\mu_{i,l}\},\{\theta_i\},\{\kappa_l\},\{\delta_{i,l}\}) \Bigg\}.$$

The necessity of additional terms in PMED-Hinge are discussed in Appendix B. The following theorem, whose proof is in Appendix G, derives a regret upper bound that matches the lower bound in Theorem 2.

Theorem 6. (Asymptotic optimality of PMED-Hinge) *Let the solution of the optimal exploration* $\mathcal{R}^{*,\mathrm{H}}_{1,\dots,L}(\{\mu_{i,l}\},\{\theta_i\},\{\kappa_l\},\{\delta_{i,l}\})$ *restricted to* $l \leq L$ *is unique at* $(\{\mu^*_{i,l}\},\{\theta^*_i\},\{\kappa^*_l\},\{0\})$. *For any* $\alpha > 0$, $\beta > 0$, *and* $\gamma > 0$, *the regret of PMED-Hinge is bounded as:*

$$\mathbb{E}[\mathrm{Reg}(T)] \leq C^*_{1,\dots,L}(\{\mu^*_{i,l}\},\{\theta^*_i\},\{\kappa^*_l\})\log T + o(\log T).$$

Note that, the assumption on the uniqueness of the solution in Theorem 6 is required to achieve an optimal coefficient on the $\log T$ factor. It is not very difficult to derive an $O(\log T)$ regret even though the uniqueness condition is not satisfied. Although our regret bound is not finite-time, the only asymptotic analysis comes from the optimal constant on the top of $\log T$ term (Lemma 11 in Appendix) and it is not very hard to derive an $O(\log T)$ finite-time regret bound.

8 Conclusion

By providing a regret lower bound and an algorithm with a matching regret bound, we gave the first complete characterization of a position-based multiple-play multi-armed bandit problem where the quality of the arms and the discount factor of the slots are unknown. We provided a way to compute the optimization problems related to the algorithm, which is of its own interest and is potentially applicable to other bandit problems.

Acknowledgements

References

[1] D. Agarwal, B.-C. Chen, and P. Elango. Spatio-temporal models for estimating click-through rate. In *WWW*, pages 21–30, 2009.

[2] V. Anantharam, P. Varaiya, and J. Walrand. Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-part i: I.i.d. rewards. *Automatic Control, IEEE Transactions on*, 32(11):968–976, 1987.

[3] P. Auer, Y. Freund, and R. E. Schapire. The non-stochastic multi-armed bandit problem. *Siam Journal on Computing*, 2002.

[4] G. Bartók, D. P. Foster, D. Pál, A. Rakhlin, and C. Szepesvári. Partial monitoring - classification, regret bounds, and algorithms. *Math. Oper. Res.*, 39(4):967–997, 2014.

[5] S. Bubeck. *Bandits Games and Clustering Foundations*. Theses, Université des Sciences et Technologie de Lille - Lille I, June 2010.

[6] A. Burnetas and M. Katehakis. Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, 17(2):122–142, 1996.

[7] A. Burnetas and M. Katehakis. Optimal adaptive policies for markov decision processes. *Math. Oper. Res.*, 22(1):222–255, Feb. 1997.

[8] R. Combes, S. Magureanu, A. Proutière, and C. Laroche. Learning to rank: Regret lower bounds and efficient algorithms. In *Proceedings of the 2015 ACM SIGMETRICS*, pages 231–244, 2015.

[9] R. Combes, M. S. Talebi, A. Proutière, and M. Lelarge. Combinatorial bandits revisited. In *NIPS*, pages 2116–2124, 2015.

[10] N. Craswell, O. Zoeter, M. J. Taylor, and B. Ramsey. An experimental comparison of click position-bias models. In *WSDM*, pages 87–94, 2008.

[11] A. Garivier and O. Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *COLT*, pages 359–376, 2011.

[12] A. Garivier and E. Kaufmann. Optimal best arm identification with fixed confidence. In *COLT*, pages 998–1027, 2016.

[13] T. L. Graves and T. L. Lai. Asymptotically efficient adaptive choice of control laws in controlled markov chains. *SIAM Journal on Control and Optimization*, 35(3):715–743, 1997.

[14] F. Guo, C. Liu, and Y. M. Wang. Efficient multiple-click models in web search. In *WSDM*, pages 124–131, 2009.

[15] P. Hall. On representatives of subsets. *Journal of the London Mathematical Society*, s1-10(1):26–30, 1935.

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

[17] J. Honda and A. Takemura. An Asymptotically Optimal Bandit Algorithm for Bounded Support Models. In *COLT*, pages 67–79, 2010.

[18] J. E. Hopcroft and R. M. Karp. An $n^{5/2}$ algorithm for maximum matchings in bipartite graphs. *SIAM Journal on Computing*, 2(4):225–231, 1973.

[19] Interactive Advertising Bureau. IAB internet advertising revenue report - 2016 full year results., 2017.

[20] S. Kale, L. Reyzin, and R. E. Schapire. Non-stochastic bandit slate problems. In *NIPS*, pages 1054–1062, 2010.

[21] S. Katariya, B. Kveton, C. Szepesvári, and Z. Wen. DCM bandits: Learning to rank with multiple clicks. In *ICML*, pages 1215–1224, 2016.

[22] KDD cup 2012 track 2, 2012.

[23] D. Kempe and M. Mahdian. A cascade model for externalities in sponsored search. In *WINE*, pages 585–596, 2008.

[24] J. Komiyama, J. Honda, and H. Nakagawa. Optimal regret analysis of thompson sampling in stochastic multi-armed bandit problem with multiple plays. In *ICML*, pages 1152–1161, 2015.

[25] J. Komiyama, J. Honda, and H. Nakagawa. Regret lower bound and optimal algorithm in finite stochastic partial monitoring. In *NIPS*, pages 1792–1800, 2015.

[26] B. Kveton, C. Szepesvári, Z. Wen, and A. Ashkan. Cascading bandits: Learning to rank in the cascade model. In *ICML*, pages 767–776, 2015.

[27] P. Lagrée, C. Vernade, and O. Cappé. Multiple-play bandits in the position-based model. In *NIPS*, pages 1597–1605, 2016.

[28] T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

[29] A. Mutapcic and S. P. Boyd. Cutting-set methods for robust convex optimization with pessimizing oracles. *Optimization Methods and Software*, 24(3):381–406, 2009.

[30] J. Nocedal and S. Wright. *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering. Springer New York, 2nd edition, 2006.

[31] A. Piccolboni and C. Schindelhauer. Discrete prediction games with arbitrary feedback and loss. In *COLT 2001 and EuroCOLT 2001*, pages 208–223, 2001.

[32] W. R. Thompson. On The Likelihood That One Unknown Probability Exceeds Another In View Of The Evidence Of Two Samples. *Biometrika*, 25:285–294, 1933.

[33] H. P. Vanchinathan, G. Bartók, and A. Krause. Efficient partial monitoring with prior information. In *NIPS*, pages 1691–1699, 2014.

[34] S. Yuan, J. Wang, and X. Zhao. Real-time bidding for online advertising: Measurement and analysis. In *Proceedings of the Seventh International Workshop on Data Mining for Online Advertising*, ADKDD '13, pages 3:1–3:8. ACM, 2013.

2	0	0	1
0	2	0	1
0	0	2	1
1	1	1	0

Figure 3: A $\tilde{N}$ matrix with $K = 4$. Greedily choosing $(1, 1), (2, 2)$, and $(3, 3)$ (of value 2 each) entries fails to find a maximal matching.

A Case in which Greedy Matching Fails

Figure 3 shows a case in which a naive greedy algorithm fails to find a maximal (perfect) matching. Consider the naive greedy algorithm that chooses the largest $\tilde{N}_{i,l}$ among all entries and tries to create a matching iteratively. Such an algorithm tries to focus on the first three diagonal elements of value 2, after that it stucks, not to yield a perfect matching.

B Discussions on Hyperparameters

The hinged version of the algorithm involves three hyperparameters: α, β, and γ. While α is necessary for assuring the quality of the estimators, we conjecture that the terms β and γ are just theoretical artifacts and unnecessary: β is for stopping the solution $\{q_{i,l}\}$ from diverging (Lemma 22 in the Appendix), which is unlikely to occur for a long time. γ is for avoiding a large value of $d_{\mathrm{KL}}(\hat{\mu}_{i,l}(t), \hat{\theta}_i(t)\hat{\kappa}_l(t))$: large deviation principle states that $d_{\mathrm{KL}}(\hat{\mu}_{i,l}(t), \theta_i^* \kappa_l^*) \sim \Theta(1/N_{i,l}(t))$, and thus MLE ($\{\theta_i^*\}, \{\kappa_l^*\}$) is unlikely to behave badly with a moderate value of $N_{i,l}(t)$.

C Parameters from KDD Cup 2012 Dataset

Table 1 shows the parameters estimated from the KDD cup 2012 dataset.

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 7. (The Chernoff bound)
Let $X_1, \ldots, X_n$ 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

Table 1: Values of $\{\theta_i^*\}$ and $\{\kappa_l^*\}$ estimated from the KDD cup 2012 dataset.

K	L	$\{\theta_i^*\}$	$\{\kappa_l^*\}$
5	3	$\{0.0463, 0.0135, 0.0127, 0.0106, 0.00629\}$	$\{1, 0.49, 0.375\}$
5	3	$\{0.0435, 0.0418, 0.0132, 0.00684, 0.00572\}$	$\{1, 0.377, 0.187\}$
6	3	$\{0.0315, 0.0208, 0.0193, 0.0182, 0.0179, 0.0177\}$	$\{1, 0.534, 0.457\}$
7	3	$\{0.0405, 0.038, 0.0265, 0.0261, 0.0256, 0.0164, 0.0112\}$	$\{1, 0.521, 0.46\}$
10	3	$\{0.0774, 0.0709, 0.0669, 0.0631, 0.043,$ $0.0393, 0.0296, 0.0217, 0.00797, 0.00219\}$	$\{1, 0.442, 0.285\}$
5	3	$\{0.0654, 0.0496, 0.0395, 0.0247, 0.0231\}$	$\{1, 0.359, 0.25\}$
8	3	$\{0.037, 0.0275, 0.0266, 0.0266, 0.0231, 0.0192, 0.0143, 0.0107\}$	$\{1, 0.542, 0.393\}$
5	3	$\{0.147, 0.0343, 0.0272, 0.0222, 0.0166, 0.0162, 0.00966\}$	$\{1, 0.624, 0.482\}$

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

E Proof of Theorem 4

Proof of Theorem 4. Consider the bipartite graph that we described in Section 4.2. By assumption, the sums of every row and column are identical, and let r_{sum} be that value. Let I and L be the left and right nodes, respectively. From Hall's marriage theorem [15], a bipartite graph has a perfect matching iff
$$|I^{\mathrm{sub}}| \leq |N_G(I^{\mathrm{sub}})| \tag{3}$$
for every subset I^{sub} of I, where $N_G(x)$ is the neighbors of x. We prove inequality (3) by contradiction. Assume that there exist an I^{sub} such that $|I^{\mathrm{sub}}| > |N_G(I^{\mathrm{sub}})|$. Then,
$$\frac{1}{N_G(I^{\mathrm{sub}})} \sum_{l \in N_G(I^{\mathrm{sub}})} \sum_{i \in [K]} r_{i,l} \geq \frac{1}{N_G(I^{\mathrm{sub}})} \sum_{i \in I^{\mathrm{sub}}} \sum_{l \in [K]} r_{i,l} = \frac{|I^{\mathrm{sub}}|}{|N_G(I^{\mathrm{sub}})|} r_{\mathrm{sum}} > r_{\mathrm{sum}}.$$
and thus at least one of the columns has a sum larger than r_{sum}, which contradicts the fact that the sum of every row and column is r_{sum}. $\qquad\square$

F Proofs of Regret Lower Bound

In this section, we prove Lemma 1 and Theorem 2. In the following, 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. The technique here is inspired from Theorem 1 in Lai and Robbins [28]. Let $\{\theta_i'\}, \{\kappa_l'\}$ be another set of parameters such that $\theta_i^* \kappa_i^* = \theta_i' \kappa_i'$ and there exists $i \neq j, i \in [L], j \in [K]$ such that $(\theta_i^* - \theta_j^*)(\theta_i' - \theta_j') < 0$. Let $\mu_{i,l}^* = \theta_i^* \kappa_l^*$ and $\mu_{i,l}' = \theta_i' \kappa_l'$. Let $(i)' \in [K]$ be the index of i-th best parameters among $\{\theta_i'\}$. With these parameters, there exists $i \in [L]$ such that $(i)' \neq i$ (i.e., the list of top-L arms is different from the ones of the true parameters). We consider a modified bandit problem with this parameters.

Let $x_{i,l}^m \in \{0,1\}$ is the reward of the m-th observation of arm i in slot l. Let
$$\widehat{\mathrm{KL}}_{i,l}(n) = \sum_{m=1}^n \log\left(\frac{x_{i,l}^m \mu_{i,l}^* + (1 - x_{i,l}^m)(1 - \mu_{i,l}^*)}{x_{i,l}^m \mu_{i,l}' + (1 - x_{i,l}^m)(1 - \mu_{i,l}')}\right),$$
and $\widehat{\mathrm{KL}} = \sum_{(i,l) \in [K] \times [L]} \widehat{\mathrm{KL}}_{i,l}(N_{i,l}(t))$. Let $\mathbb{P}'$ and $\mathbb{E}'$ be the probability and the expectation with respect to the modified game, respectively. Then, for any event $\mathcal{E}$,
$$\mathbb{P}'[\mathcal{E}] = \mathbb{E}\left[\mathbf{1}[\mathcal{E}] \exp\left(-\widehat{\mathrm{KL}}\right)\right] \tag{4}$$
holds. Now, let us define the following events:
$$\mathcal{D}_1 = \left\{\sum_{(i,l) \in [K] \times [L]} N_{i,l}(t) d_{\mathrm{KL}}(\mu_{i,l}^*, \mu_{i,l}') < (1 - \epsilon)\log T, \bigcap_{i \in [L]: i \neq (i)'} N_{(i)',i}(t) < \sqrt{T}\right\},$$
$$\mathcal{D}_2 = \left\{\widehat{\mathrm{KL}} \leq \left(1 - \frac{\epsilon}{2}\right)\log T\right\},$$
$$\mathcal{D}_{12} = \mathcal{D}_1 \cap \mathcal{D}_2,$$
$$\mathcal{D}_{1\backslash 2} = \mathcal{D}_1 \cap \mathcal{D}_2^c.$$

Firstly, we show $\mathbb{P}[\mathcal{D}_{12}] = o(1)$. From (4),

$$\mathbb{P}'[\mathcal{D}_{12}] \geq \mathbb{E}\left[\mathbf{1}[\mathcal{D}_{12}] \exp\left(-\left(1 - \frac{\epsilon}{2}\right)\log T\right)\right] = T^{-(1-\epsilon/2)}\mathbb{P}[\mathcal{D}_{12}].$$

By using this we have

$$\mathbb{P}[\mathcal{D}_{12}] \leq T^{(1-\epsilon/2)}\mathbb{P}'[\mathcal{D}_{12}]$$

$$\leq T^{(1-\epsilon/2)}\mathbb{P}'\left[\bigcap_{i\in[L]:i\neq(i)'} \{N_{(i)',i}(t) < \sqrt{T}\}\right]$$

$$\leq T^{(1-\epsilon/2)} \sum_{i\in[L]:i\neq(i)'} \mathbb{P}'\left[T - N_{(i)',i}(t) > T - \sqrt{T}\right]$$

$$\leq T^{(1-\epsilon/2)} \sum_{i\in[L]:i\neq(i)'} \frac{\mathbb{E}'[T - N_{(i)',i}(t)]}{T - \sqrt{T}} \qquad \text{(by the Markov inequality).} \qquad (5)$$

Since this algorithm is strongly consistent, $\mathbb{E}'[T - N_{(i)',i}(t)] \to o(T^a)$ for any $a > 0$. Therefore, the RHS of the last line of (5) is $o(T^{a-\epsilon/2})$, which, by choosing sufficiently small a, converges to zero as $T \to \infty$. In summary, $\mathbb{P}[\mathcal{D}_{12}] = o(1)$.

Secondly, we show $\mathbb{P}[\mathcal{D}_{1\setminus 2}] = o(1)$. We have

$$\mathbb{P}[\mathcal{D}_{1\setminus 2}] = \mathbb{P}\left[\sum_{(i,l)\in[K]\times[L]} N_{i,l}(t)d_{\mathrm{KL}}(\mu_{i,l}, \mu'_{i,l}) < (1-\epsilon)\log T,\right.$$

$$\left. \bigcap_{i\in[L]:i\neq(i)'} N_{(i)',i}(t) < \sqrt{T}, \sum_{(i,l)\in[K]\times[L]} \widehat{\mathrm{KL}}_i(N_{i,l}(t)) > \left(1 - \frac{\epsilon}{2}\right)\log T\right].$$

Note that

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

is the maximum of the sum of positive-mean random variables, and thus converges to is average (c.f., Bubeck [5, Lemma 10.5]). Namely,

$$\lim_{N\to\infty} \max_{1\leq n\leq N} \frac{\widehat{\mathrm{KL}}_{i,l}(n)}{N} = d_{\mathrm{KL}}(\mu^*_{i,l}, \mu'_{i,l})$$

almost surely. Therefore,

$$\lim_{T\to\infty} \frac{\max_{\{N_{i,l}(t)\}\in\mathbb{N}^N, \sum_{(i,l)\in[K]\times[L]} N_{i,l}(t)d_{\mathrm{KL}}(\mu^*_{i,l},\mu'_{i,l})<(1-\epsilon)\log T} \sum_{(i,l)\in[K]\times[L]} \widehat{\mathrm{KL}}_i(N_{i,l}(t))}{\log T} = 1-\epsilon$$

almost surely. By using this fact and $1 - \epsilon/2 > 1 - \epsilon$, we have

$$\mathbb{P}\left[\max_{\{N_{i,l}(t)\}\in\mathbb{N}^{KL}, \sum_{(i,l)\in[K]\times[L]} N_{i,l}(t)d_{\mathrm{KL}}(\mu^*_{i,l},\mu'_{i,l})<(1-\epsilon)\log T} \sum_{(i,l)\in[K]\times[L]} \widehat{\mathrm{KL}}_i(N_{i,l}(t)) > \left(1 - \frac{\epsilon}{2}\right)\log T\right] = o(1).$$

In summary, we obtain $\mathbb{P}\left[\mathcal{D}_{1\setminus 2}\right] = o(1)$.

We here have

$$\mathcal{D}_1 = \left\{\sum_{(i,l)\in[K]\times[L]} N_{i,l}(t)d_{\mathrm{KL}}(\mu^*_{i,l}, \mu'_{i,l}) < (1-\epsilon)\log T\right\} \cap \left\{\bigcap_{i\in[L]:i\neq(i)'} N_{(i)',i}(t) < \sqrt{T}\right\}$$

$$\supseteq \left\{\sum_{(i,l)\in[K]\times[L]} N_{i,l}(t)d_{\mathrm{KL}}(\mu_{i,l}, \mu'_{i,l}) + \sum_{i\in[L]:i\neq(i)'} \frac{(1-\epsilon)\log T}{\sqrt{T}}N_{(i)',i}(t) < (1-\epsilon)\log T\right\},$$

14

where we used the fact that $\{A < C\} \cap \{B < C\} \supseteq \{A + B < C\}$ for $A, B > 0$ in the last line. Note that, by using the result of the previous steps, $\mathbb{P}[\mathcal{D}_1] = \mathbb{P}[\mathcal{D}_{12}] + \mathbb{P}[\mathcal{D}_{1\setminus 2}] = o(1)$. By using the complementary of this fact,

$$
\mathbb{P}\left[\sum_{(i,l)\in[K]\times[L]} N_{i,l}(t) d_{\mathrm{KL}}(\mu_{i,l}^*, \mu_{i,l}') + \sum_{i\in[L]:i\neq(i)'} \frac{(1-\epsilon)\log T}{\sqrt{T}} N_{(i)',i}(t) \geq (1-\epsilon)\log T \right]
$$
$$
\geq \mathbb{P}[\mathcal{D}_1^c] = 1 - o(1).
$$

Using the Markov inequality yields

$$
\mathbb{E}\left[\sum_{(i,l)\in[K]\times[L]} N_{i,l}(t) d_{\mathrm{KL}}(\mu_{i,l}^*, \mu_{i,l}') + \frac{(1-\epsilon)\log T}{\sqrt{T}} \sum_{i\in[L]:i\neq(i)'} N_{(i)',i}(t) \right] \geq (1-\epsilon)(1-o(1))\log T.
$$
(6)

Because $\mathbb{E}[N_{(i)',i}(t)]$ is subpolynomial as a function of T due to the consistency, the second term in LHS of (6) is $o(1)$ and thus negligible. Lemma 1 follows from the fact that (6) holds for sufficiently small ϵ and arbitrary $\{\theta_i', \kappa_l'\}$. $\qquad\square$

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
$$
\mathbb{E}[\mathrm{Reg}(T_s)] < (1-\delta) C_{1,\ldots,L}^*(\{\mu_{i,l}^*\}, \{\theta_i^*\}, \{\kappa_l^*\}) \log T_s ,
$$

that is,

$$
\sum_{(i,l)\in[K]\times[L]} \Delta_{i,l} \frac{\mathbb{E}[N_{i,l}(T_s)]}{(1-\delta)\log T_s} < C_{1,\ldots,L}^*(\{\mu_{i,l}^*\}, \{\theta_i^*\}, \{\kappa_l^*\}) .
$$

From the definition of $C_{1,\ldots,L}^*$, there exists $\{\theta_i^s\} \in \mathcal{T}_{1,\ldots,L}^c$ such that

$$
\sum_{(i,l)\in[K]\times[L]} d_{\mathrm{KL}}\left(\theta_i^* \kappa_l^*, \theta_i^s \frac{\theta_l^* \kappa_l^*}{\theta_l^s} \right) \frac{\mathbb{E}[N_{i,l}(T_s)]}{(1-\delta)\log T_s} < 1
$$

Since $\mathcal{T}_{1,\ldots,L}^c$ is compact, there exists a subsequence $s_0 < s_1 < \cdots$ such that $\lim_{u\to\infty}\{\theta_i^{s_u}\} = \{\theta_i\}$ for some $\{\theta_i\} \in \mathcal{T}_{1,\ldots,L}^c$. Therefore from the lower semicontinuity of the divergence we obtain

$$
1 \geq \sum_{(i,l)\in[K]\times[L]} \liminf_{u\to\infty} \frac{\mathbb{E}[N_i(T_{s_u})]}{(1-\delta)\log T_{s_u}} d_{\mathrm{KL}}\left(\theta_i^* \kappa_l^*, \theta_i^{s_u} \frac{\theta_l^* \kappa_l^*}{\theta_l^{s_u}} \right)
$$
$$
\geq \sum_{(i,l)\in[K]\times[L]} \liminf_{s\to\infty} \frac{\mathbb{E}[N_i(T_s)]}{(1-\delta)\log T_s} d_{\mathrm{KL}}\left(\theta_i^* \kappa_l^*, \theta_i \frac{\theta_l^* \kappa_l^*}{\theta_l} \right) ,
$$

which contradicts Lemma 1.

$\qquad\square$

G   Main Regret Bound: Proof of Theorem 6

In this section, we prove the asymptotic optimality of PMED-Hinge. First, we define the following events that are important in bounding regret. Let $\hat{i}(t)$ be the i-th largest arm based on $\hat{\theta}_i(t)$ (ties are broken arbitrarily).
$$
\mathcal{W}(t) = \bigcup_{i\in[L]} \left\{ \hat{i}(t) \neq i \right\}
$$

Henceforth, we abbreviate the event that some L-allocation that includes $(i,l) \in [K] \times [L]$ is put into L_N to "pair (i,l) is put into L_N". Let $\mathcal{J}_{i,l}(t)$ be the event that pair (i,l) is put into L_N at round

15

t. Let $\mathcal{X}(t)$ is the event that at least one arm is put into L_N before Line 18 in Algorithm 1. Namely,

$$
\mathcal{X}(t) = \Bigg\{ \bigcup_{(i,l) \in [K] \times [L]} \{N_{i,l}(t) < \alpha \sqrt{\log t}\} \cup \bigcup_{(i \neq j) \in [K] \times [K]} \{|\hat{\theta}_i(t) - \hat{\theta}_j(t)| < \beta/(\log \log t)\}
$$

$$
\cup \bigcup_{(l \neq m) \in [K] \times [K]} \{|\hat{\kappa}_l(t) - \hat{\kappa}_m(t)| < \beta/(\log \log t)\} \cup \bigcup_{(i,l) \in [K] \times [L]} \Big\{ d_{\mathrm{KL}}(\hat{\mu}_{i,l}(t), \hat{\theta}_i(t)\hat{\kappa}_l(t)) \geq f(N_{i,l}(t)) \Big\} \Bigg\}.
$$

Let $\mathcal{Y}_{i,l}(t)$ be the event that pair (i, l) is put into L_N in Line 18 of Algorithm 1. Moreover, let

$$
\mathcal{Z}_\delta(t) = \bigcap_{i \in [K]} \{|\hat{\theta}_i(t) - \theta_i^*| < \delta\} \cap \bigcap_{l \in [L]} \{|\hat{\kappa}_l(t) - \kappa_l^*| < \delta\}.
$$

That is, the estimator $\{\hat{\theta}_i(t), \hat{\kappa}_l(t)\}$ is sufficiently close to the set of true values.

By using $\Delta_{i,l} \leq 1$, the regret is decomposed into the following terms:

$$
\mathrm{Reg}(T) = \sum_{(i,l) \in [K] \times [L]} \Delta_{i,l} \sum_{t=1}^{T} \mathbf{1}[I_i(t) = l]
$$

$$
\leq \sum_{(i,l) \in [K] \times [L]} \Delta_{i,l} \sum_{t=1}^{T} \mathbf{1}[\mathcal{J}_{i,l}(t)] + K^2
$$

$$
\leq K \sum_{t=1}^{T} \mathbf{1}[\mathcal{X}(t), \mathcal{J}_{i,l}(t)] + \Delta_{i,l} \sum_{t=1}^{T} \mathbf{1}[\mathcal{X}^c(t), \mathcal{Y}_{i,l}(t)] + K^2
$$

$$
\leq K \sum_{t=1}^{T} \mathbf{1}[\mathcal{X}(t), \mathcal{J}_{i,l}(t)] + K \sum_{t=1}^{T} \mathbf{1}[\mathcal{X}^c(t), \mathcal{W}(t)] + \sum_{(i,l) \in [K] \times [L]} \Delta_{i,l} \sum_{t=1}^{T} \mathbf{1}[\mathcal{W}^c(t), \mathcal{Y}_{i,l}(t), \mathcal{Z}_\delta(t)]
$$

$$
+ \sum_{(i,l) \in [K] \times [L]} \sum_{t=1}^{T} \mathbf{1}[\mathcal{X}^c(t), \mathcal{W}^c(t), \mathcal{Y}_{i,l}(t), \mathcal{Z}_\delta^c(t)] + K^2 \tag{7}
$$

The following lemmas bound each term of (7), and combining them completes the proof.

Lemma 9. (Sublog exploration) *The following inequality holds:*

$$
\sum_{t=1}^{T} \sum_{(i,l) \in [K] \times [L]} \mathbb{P}\left[\mathcal{X}(t), \mathcal{J}_{i,l}(t)\right] = o(\log T)
$$

Lemma 10. (Misidentification) *The following inequality holds:*

$$
\sum_{t=1}^{T} \mathbb{P}\left[\mathcal{X}^c(t), \mathcal{W}(t)\right] = O(1)
$$

Lemma 11. (Leading term) *There exists a continuous function $\epsilon(\delta)$ such that $\epsilon(\delta) \to 0$ as $\delta \to 0$, and the following inequality holds:*

$$
\sum_{t=1}^{T} \mathbf{1}\left[\mathcal{W}^c(t), \mathcal{Y}_{i,l}(t), \mathcal{Z}_\delta(t)\right] = (1 + \epsilon(\delta))R_{i,l}^* \log T + 1.
$$

where $R_{i,l}^*$ is the (i, l) entry of the optimal solution $\mathcal{R}_{1,\ldots,L}^*(()\{\mu_{i,l}^*\}, \{\theta_i^*\}, \{\kappa_i^*\})$.

Lemma 12. (Case in which estimation quality is low) *The following inequality holds:*

$$
\sum_{t=1}^{T} \mathbb{P}\left[\mathcal{X}^c(t), \mathcal{W}^c(t), \mathcal{Y}_{i,l}(t), \mathcal{Z}_\delta^c(t)\right] = o(\log T).
$$

The following sections prove Lemmas 9–12. We use $\hat{\mu}_{i,l}^n$ to denote $\hat{\mu}_{i,l}(t)$ when $N_{i,l}(t) = n$.

H   Proof of Lemma 9

Lemma 13. (Convergence of the hinge of MLE) *Let*

$$\mathcal{C}(t) = \bigcap_{(i,l)\in[K]\times[L]} \left\{ d_{\mathrm{KL}}(\hat{\mu}_{i,l}(t), \hat{\theta}_i(t)\hat{\kappa}_l(t)) \le f(N_{i,l}(t)) \right\}. \tag{8}$$

The following inequality holds of any $(i,l) \in [K] \times [L]$:

$$\sum_{t=1}^{T} \mathbb{P}\left[\mathcal{J}_{i,l}(t), \mathcal{C}^c(t)\right] = O(1).$$

Proof of Lemma 13. Let $N^C(t)$ be the number of rounds before t such that pair (i,l) is put into L_N by Line 10 of Algorithm 1. Note that $\{\mathcal{J}_{i,l}(t), \mathcal{C}^c(t), N^C(t) = n\}$ occurs at most twice because if $\mathcal{C}^c(t)$ occurs then (i,l) is put into L_N by Line 10. By using this fact, we obtain

$$\sum_{t=1}^{T} \mathbb{P}\left[\mathcal{J}_{i,l}(t), \mathcal{C}^c(t)\right]$$

$$\le 2\sum_{n=1}^{T} \mathbb{P}\left[\bigcup_{t=n}^{T} \{\mathcal{J}_{i,l}(t), \mathcal{C}^c(t), N^C(t) = n\}\right]$$

$$\le 2\sum_{n=1}^{T} \mathbb{P}\left[\bigcup_{t=n}^{T} \{\mathcal{J}_{i,l}(t), \mathcal{C}^c(t), N_{i,l}(t) \ge n\}\right]$$

$$\le 2\sum_{n=1}^{T} \sum_{(i,l)\in[K]\times[L]} \sum_{n'=n}^{T} \mathbb{P}\left[d_{\mathrm{KL}}(\hat{\mu}_{i,l}^{n'}, \theta_i^*\kappa_l^*) > f(n')\right] \le 2\sum_{n=1}^{T} \sum_{(i,l)\in[K]\times[L]} \sum_{n'=n}^{T} e^{-n'f(n')} = O(1)$$

$$\square$$

where we have used the facts that $N^C(t) = n \Rightarrow \bigcap_{(i,l)}\{N_{i,l}(t) \ge n\}$ and $\mathcal{C}^c(t)$ implies $\bigcup_{(i,l)}\{d_{\mathrm{KL}}(\hat{\mu}_{i,l}^{n'}, \theta_i^*\kappa_l^*) > f(n')\}$.

Lemma 14. *For any $i \ne j \in [K]$, the following inequality holds:*

$$\sum_{t=1}^{T} \mathbb{P}\left[\bigcap_{(i',l')\in[K]\times[L]} N_{i',l'}(t) \ge \alpha\sqrt{\log t}, \mathcal{J}_{i,l}(t), \mathcal{C}(t), |\hat{\theta}_i(t) - \hat{\theta}_j(t)| < \beta/(\log\log t)\right] = O(1).$$

Proof of Lemma 14. Let $\Delta_\theta = |\theta_i^* - \theta_j^*| > 0$. Note that

$$\left\{ |\hat{\theta}_i(t) - \hat{\theta}_j(t)| < \beta/(\log\log t), t \ge e^{e^{5\beta/\Delta_\theta}} \right\}$$

implies $|\hat{\theta}_i(t) - \hat{\theta}_j(t)| \le \Delta_\theta/5$. Moreover,

$$\left\{ N_{i',l'}(t) > \alpha\sqrt{\log t}, t > e^{\left(\frac{(5^2\gamma)^2}{4\alpha\Delta_\theta^2}\right)^2} \right\}$$

implies that $2(\Delta_\theta/5)^2 > f(N_{i',l'}(t))$. Let $N^D(t)$ be the number of rounds before t such that the arms are put into L_N by Line 9. Then $\{\mathcal{J}_{i,l}(t), N^D(t) = n\}$ occurs at most twice. By using these,

17

we obtain

$$\sum_{t=1}^{T}\mathbb{P}\left[\bigcap_{(i',l')\in[K]\times[L]}N_{i',l'}(t)>\alpha\sqrt{\log t},|\hat{\theta}_i(t)-\hat{\theta}_j(t)|<\beta/(\log\log t),\mathcal{J}_{i,l}(t),\mathcal{C}(t)\right]$$

$$\leq\sum_{n=1}^{T}\mathbb{P}\left[\bigcup_{t=1}^{T}\left\{\bigcap_{(i',l')\in[K]\times[L]}N_{i',l'}(t)>\alpha\sqrt{\log t},|\hat{\theta}_i(t)-\hat{\theta}_j(t)|<\beta/(\log\log t),N^D(t)=n\right\},\mathcal{J}_{i,l}(t),\mathcal{C}(t)\right]$$

$$\leq\max(e^{\left(\frac{(5^2\gamma)^2}{4\alpha\Delta_\theta^2}\right)^2},e^{e^{5\beta/\Delta_\theta}})+\sum_{n=1}^{T}\mathbb{P}\left[\bigcup_{t=1}^{T}\left\{\mathcal{C}(t),|\hat{\theta}_i(t)-\hat{\theta}_j(t)|<\Delta_\theta/5,2(\Delta_\theta/5)^2>f(N_{i,l}(t)),N^D(t)=n\right\}\right]$$

$$\leq O(1)+\sum_{n=1}^{T}\mathbb{P}\left[\bigcup_{t=1}^{T}\left\{d_{\mathrm{KL}}(\hat{\mu}_{i,1}(t),\hat{\theta}_i(t))<2(\Delta_\theta/5)^2,d_{\mathrm{KL}}(\hat{\mu}_{j,1}(t),\hat{\theta}_j(t))<2(\Delta_\theta/5)^2,\right.\right.$$
$$\left.\left.|\hat{\theta}_i(t)-\hat{\theta}_j(t)|<\Delta_\theta/5,N^D(t)=n\right\}\right]$$

$$\leq O(1)+\sum_{n=1}^{T}\mathbb{P}\left[\bigcup_{t=1}^{T}\left\{(\hat{\mu}_{i,1}(t)-\hat{\theta}_i(t))^2<(\Delta_\theta/5)^2,(\hat{\mu}_{j,1}(t)-\hat{\theta}_j(t))^2<(\Delta_\theta/5)^2,\right.\right.$$
$$\left.\left.|\hat{\theta}_i(t)-\hat{\theta}_j(t)|<\Delta_\theta/5,N^D(t)=n\right\}\right]$$

(by Pinsker's inequality)

$$\leq O(1)+\sum_{n=1}^{T}\mathbb{P}\left[\bigcup_{t=1}^{T}\left\{|\hat{\mu}_{i,1}(t)-\hat{\theta}_i(t)|<\Delta_\theta/5,|\hat{\mu}_{j,1}(t)-\hat{\theta}_j(t)|<\Delta_\theta/5,|\hat{\theta}_i(t)-\hat{\theta}_j(t)|<\Delta_\theta/5,N^D(t)=n\right\}\right]$$

$$\leq O(1)+\sum_{n=1}^{T}\mathbb{P}\left[\bigcup_{t=1}^{T}\left\{|\hat{\mu}_{i,1}(t)-\theta_i^*|>\Delta_\theta/5\cup|\hat{\mu}_{j,1}(t)-\theta_j^*|>\Delta_\theta/5\right\}\right]$$

(by the triangular inequality and the definition of Δ_θ.)

$$\leq O(1)+\sum_{n=1}^{T}\mathbb{P}\left[\bigcup_{t=1}^{T}\left\{|\hat{\mu}_{i,1}(t)-\theta_i^*|>\Delta_\theta/5\right\}\right]+\sum_{n=1}^{T}\mathbb{P}\left[\bigcup_{t=1}^{T}\left\{|\hat{\mu}_{j,1}(t)-\theta_j^*|>\Delta_\theta/5\right\}\right].\quad(9)$$

Here,

$$\sum_{n=1}^{T}\mathbb{P}\left[\bigcup_{t=1}^{T}\left\{|\hat{\mu}_{i,1}(t)-\theta_i^*|>\Delta_\theta/5,N^D(t)=n\right\}\right]$$
$$\leq\sum_{n=1}^{T}\sum_{n'=n}^{T}\mathbb{P}\left[|\hat{\mu}_{i,1}^{n'}-\theta_i^*|>\Delta_\theta/5\right]$$
(by the fact that $N^D(t)=n$ implies $N_{i,l}(t)\geq n$.)
$$\leq\sum_{n=1}^{T}\sum_{n'=n}^{T}e^{2n'(\Delta_\theta/5)^2}=O(1),$$
(by Chernoff bound and Pinsker's inequality)

$$(10)$$

and the same inequality also holds for j, which completes the proof. $\qquad\square$

Lemma 15. *For any $l\neq m\in[K]$, the following inequality holds:*

$$\sum_{t=1}^{T}\mathbb{P}\left[\bigcap_{(i',l')\in[K]\times[L]}N_{i',l'}(t)\geq\alpha\sqrt{\log t},|\hat{\kappa}_l(t)-\hat{\kappa}_m(t)|<\beta/(\log\log t),\mathcal{J}_{i,l}(t),\mathcal{C}(t)\right]=O(1).$$

Proof of Lemma 15. Let $\Delta_\kappa = \min_{l \neq m} \theta_1 |\kappa_l^* - \kappa_m^*|$. Similar to Lemma 14, we have

$$\sum_{t=1}^{T} \mathbb{P}\left[\bigcap_{(i',l') \in [K] \times [L]} N_{i',l'}(t) \geq \alpha \sqrt{\log t}, |\hat{\kappa}_l(t) - \hat{\kappa}_m(t)| < \beta/(\log \log t), \mathcal{J}_{i,l}(t), \mathcal{C}(t)\right]$$

$$\leq \sum_{n=1}^{T} \mathbb{P}\left[\bigcup_{t=1}^{T}\left\{\bigcap_{(i',l') \in [K] \times [L]} N_{i',l'}(t) > \alpha \sqrt{\log t}, |\hat{\kappa}_l(t) - \hat{\kappa}_m(t)| < \beta/(\log \log t), N^D(t) = n\right\}, \mathcal{J}_{i,l}(t), \mathcal{C}(t)\right]$$

$$\leq \max(e^{\left(\frac{(5^2 \gamma)^2}{4\alpha\Delta_\kappa^2}\right)^2}, e^{e^{5\beta/\Delta_\kappa}}) + \sum_{n=1}^{T} \mathbb{P}\left[\bigcup_{t=1}^{T}\left\{\mathcal{C}(t), |\hat{\kappa}_l(t) - \hat{\kappa}_m(t)| < \Delta_\kappa/5, 2(\Delta_\kappa/5)^2 > f(N_{i,l}(t)), N^D(t) = n\right\}\right]$$

$$\leq O(1) + \sum_{n=1}^{T} \mathbb{P}\left[\bigcup_{t=1}^{T}\left\{d_{\mathrm{KL}}(\hat{\mu}_{1,l}(t), \hat{\theta}_1(t)\hat{\kappa}_l(t)) < 2(\Delta_\kappa/5)^2, d_{\mathrm{KL}}(\hat{\mu}_{1,m}(t), \hat{\theta}_1(t)\hat{\kappa}_m(t)) < 2(\Delta_\kappa/5)^2,\right.\right.$$

$$\left.\left.|\hat{\kappa}_l(t) - \hat{\kappa}_m(t)| < \Delta_\kappa/5, N^D(t) = n\right\}\right]$$

$$\leq O(1) + \sum_{n=1}^{T} \mathbb{P}\left[\bigcup_{t=1}^{T}\left\{(\hat{\mu}_{1,l}(t) - \hat{\theta}_1(t)\hat{\kappa}_l(t))^2 < (\Delta_\kappa/5)^2, (\hat{\mu}_{1,m}(t) - \hat{\theta}_1(t)\hat{\kappa}_m(t))^2 < (\Delta_\kappa/5)^2,\right.\right.$$

$$\left.\left.|\hat{\kappa}_l(t) - \hat{\kappa}_m(t)| < \Delta_\kappa/5, N^D(t) = n\right\}\right]$$

$$\leq O(1) + \sum_{n=1}^{T} \mathbb{P}\left[\bigcup_{t=1}^{T}\left\{|\hat{\mu}_{1,l}(t) - \hat{\theta}_1(t)\hat{\kappa}_l(t)| < \Delta_\kappa/5,\right.\right.$$

$$\left.\left.|\hat{\mu}_{1,m}(t) - \hat{\theta}_1(t)\hat{\kappa}_m(t)| < \Delta_\kappa/5, \theta_1|\hat{\kappa}_l(t) - \hat{\kappa}_m(t)| < \Delta_\kappa/5, N^D(t) = n\right\}\right]$$

$$\leq O(1) + \sum_{n=1}^{T} \mathbb{P}\left[\bigcup_{t=1}^{T}\left\{|\hat{\mu}_{1,l}(t) - \mu_{1,l}^*| > \Delta_\kappa/5 \cup |\hat{\mu}_{1,m}(t) - \mu_{1,m}^*| > \Delta_\kappa/5\right\}\right]$$

$$\leq O(1) + \sum_{n=1}^{T} \mathbb{P}\left[\bigcup_{t=1}^{T}\left\{|\hat{\mu}_{1,l}(t) - \mu_{1,l}^*| > \Delta_\kappa/5\right\}\right] + \sum_{n=1}^{T} \mathbb{P}\left[\bigcup_{t=1}^{T}\left\{|\hat{\mu}_{1,m}(t) - \mu_{1,m}^*| > \Delta_\kappa/5\right\}\right].$$
(11)

Here,

$$\sum_{n=1}^{T} \mathbb{P}\left[\bigcup_{t=1}^{T}\left\{|\hat{\mu}_{1,l}(t) - \mu_{1,l}^*| > \Delta_\kappa/5, N^D(t) = n\right\}\right]$$

$$\leq \sum_{n=1}^{T} \sum_{n'=n}^{T} \mathbb{P}\left[|\hat{\mu}_{1,l}^{n'} - \mu_{1,l}^*| > \Delta_\kappa/5\right]$$

$$\leq \sum_{n=1}^{T} \sum_{n'=n}^{T} e^{2n'(\Delta_\kappa/5)^2} = O(1),$$
(12)

and the same inequality also holds for m, which completes the proof. $\square$

Proof of 9. The event

$$\{N_{i',l'}(t) < \alpha \sqrt{\log t}, \mathcal{J}_{i',l'}(t), N_{i',l'}(t) = n\}$$

19

occurs at most twice because if it occurs then pair (i', l') is put into L_N immediately. Taking these into consideration,

$$\sum_{t=1}^{T} \mathbb{P}\left[\mathcal{X}(t), \mathcal{J}_{i,l}(t)\right]$$

$$\leq 2\sum_{n=1}^{T} \mathbb{P}\left[\bigcup_{t=n}\left\{\bigcup_{(i',l')\in[K]\times[L]} N_{i',l'}(t) < \alpha\sqrt{\log T}, N_{i',l'}(t) = n\right\}\right] + \sum_{t=1}^{T} \mathbb{P}\left[\mathcal{J}_{i,l}(t), \mathcal{C}^c(t)\right]$$

$$+ \sum_{t=1}^{T}\sum_{i\neq j} \mathbb{P}\left[\bigcap_{(i',l')\in[K]\times[L]} N_{i',l'}(t) \geq \alpha\sqrt{\log t}, |\hat{\theta}_i(t) - \hat{\theta}_j(t)| < \beta/(\log\log t), \mathcal{J}_{i,l}(t), \mathcal{C}(t)\right]$$

$$+ \sum_{t=1}^{T}\sum_{l\neq m} \mathbb{P}\left[\bigcap_{(i',l')\in[K]\times[L]} N_{i',l'}(t) \geq \alpha\sqrt{\log t}, |\hat{\kappa}_l(t) - \hat{\kappa}_m(t)| < \beta/(\log\log t), \mathcal{J}_{i,l}(t), \mathcal{C}(t)\right]$$

$$\tag{13}$$

The first term of (13) is $O(\sqrt{\log T})$, and the second, third, and fourth terms of (13) are $O(1)$ by Lemma 13, 14, and 15 respectively. □

I Proof of Lemma 10

Remember that $\hat{1}(t), \ldots, \hat{L}(t)$ are the empirical top-L arm based on MLE. The following lemma guarantees the convergence of $\hat{\mu}_{\hat{l}(t),l}(t)$ for each $l \in [L]$.

Lemma 16. (Convergence of the estimated best arm) *Let*

$$\mathcal{P}(t) = \bigcup_{l\in[L]} \left\{d_{\mathrm{KL}}(\hat{\mu}_{\hat{1}(t),l}(t), \theta^*_{\hat{1}(t)}\kappa^*_l)\} > f(N_{\hat{1}(t),l}(t))\right\}.$$

Then, the following inequality holds:

$$\sum_{t=1}^{T} \mathbb{P}\left[\mathcal{P}(t)\right] = O(1). \tag{14}$$

Proof of Lemma 16. Let $(1), \ldots, (L)$ be L distinct elements among $[K]$. Note that,

$$\left\{\hat{1}(t) = (1), \ldots, \hat{L}(t) = (L), \bigcup_{l\in[L]} d_{\mathrm{KL}}(\hat{\mu}_{(l),l}(t), \theta^*_{(l)}\kappa^*_l)\} > f(N_{(l),l}(t)), \bigcup_{l\in[L]} N_{(l),l}(t) \leq n\right\}$$

occurs at most $K^2 n$ rounds because $((1), 1), \ldots, ((L), L)$ are put into L_N by Line 22 of Algorithm 1 under this event and $|L_N|$ is at most K^2. By using this, we have

$$\mathbb{E}\left[\hat{1}(t) = (1), \ldots, \hat{L}(t) = (L), \bigcup_{l \in [L]} d_{\mathrm{KL}}(\hat{\theta}_{(l)}(t)\hat{\kappa}_l(t), \theta^*_{(l)}\kappa^*_l)\} > f(N_{(l),l}(t))\right]$$

$$\leq \sum_{n=2}^{\infty} K^2 n \mathbb{P}\left[\bigcup_{t=1}^{T}\left(\bigcup_{l \in [L]} d_{\mathrm{KL}}(\hat{\mu}_{(l),l}(t), \theta^*_{(l)}\kappa^*_l)\} > f(N_{(l),l}(t)), \bigcup_{l \in [L]}\{N_{(l),l}(t) \leq n\} \setminus \bigcup_{l \in [L]}\{N_{(l),l}(t) \leq n-1\}\right)\right]$$

$$\leq \sum_{n=1}^{\infty} K^2 n \mathbb{P}\left[\bigcup_{t=1}^{T}\left(\bigcup_{l \in [L]} d_{\mathrm{KL}}(\hat{\mu}_{(l),l}(t), \theta^*_{(l)}\kappa^*_l)\} > f(N_{(l),l}(t)), \bigcap_{l \in [L]}\{N_{(l),l}(t) \geq n\}\right)\right]$$

$$\leq \sum_{n=1}^{\infty} K^2 n \sum_{l \in [L]} \sum_{n_l \geq n} \mathbb{P}\left[d_{\mathrm{KL}}(\hat{\mu}^{n_l}_{(l),l}, \theta^*_{(l)}\kappa^*_l)\} > f(n_l)\right]$$

$$\leq \sum_{n=1}^{\infty} 2K^2 n \sum_{l \in [L]} \sum_{n_l \geq n} e^{-n_l f(n_l)}$$

(by Chernoff bound from both sides)

$$= O(1).$$

Taking a union bound of $(1), \ldots, (L)$ over all L distinct elements among $[K]$ complete the proof. $\qquad\square$

Lemma 17. (Minimum divergence gap) *There exists a constant C_{mgap} such that*

$$\left\{N_{i,l}(t) > \alpha\sqrt{\log t}, \bigcap_{i,l}(d_{\mathrm{KL}}(\hat{\mu}_{i,l}(t), \theta^*_i\kappa^*_l) - f(N_{i,l}(t)))_+ = 0, \mathcal{W}^c(t)\right\}$$

cannot occur for $t \geq C_{\mathrm{mgap}}$.

Proof of Lemma 17. Event $\mathcal{W}^c(t)$ implies that there exists a pair $i, j \in [K]$ such that

$$\{(\theta^*_i - \hat{\theta}_i(t))(\theta^*_j - \hat{\theta}_j(t)) \leq 0\},$$

which implies $\max(|\theta^*_i - \hat{\theta}_i(t)|, |\theta^*_j - \hat{\theta}_j(t)|) > |\theta^*_i - \theta^*_j|/2$. Let $\Delta = |\theta^*_i - \theta^*_j|$. Without loss of generality we assume $|\theta^*_i - \hat{\theta}_i(t)| > \Delta/2$. Remember that $\kappa^*_1 = 1$. By using the Pinsker's inequality

$$\max(d_{\mathrm{KL}}(\hat{\mu}_{i,1}(t), \theta^*_i), d_{\mathrm{KL}}(\hat{\mu}_{i,1}(t), \hat{\theta}_i(t))) \geq \Delta^2/2 \qquad (15)$$

Note that, by the definition of MLE

$$\bigcap_{i,l}\left\{(d_{\mathrm{KL}}(\hat{\mu}_{i,l}(t), \theta^*_i\kappa^*_l) - f(N_{i,l}(t)))_+ = 0\right\} \Rightarrow \bigcap_{i,l}\left\{\left(d_{\mathrm{KL}}(\hat{\mu}_{i,l}(t), \hat{\theta}_i(t)\hat{\kappa}_l(t)) - f(N_{i,l}(t))\right)_+ = 0\right\}.$$
$$(16)$$

Inequalities (15) and (16) do not hold simultaneously for $f(N_{i,l}(t)) < \Delta^2/2$, and thus the Lemma holds with $C_{\mathrm{mgap}} = \exp(\frac{16\gamma^4}{\Delta^8\alpha^2})$ by using the condition $N_{i,j}(t) > \alpha\sqrt{\log t}$. $\qquad\square$

Proof of Lemma 10. Let

$$\mathcal{H}(t) = \{(i, l) \in [K] \times [L] : (d_{\mathrm{KL}}(\hat{\mu}_{i,l}(t), \theta^*_i\kappa^*_l) - f(N_{i,l}))_+ > 0\}.$$

Let $\mathcal{G} \in 2^{[K] \times [L]} \setminus \emptyset$ be arbitrary. Note that, if

$$\left\{\log t \geq \sum_{(i,l) \in \mathcal{G}} N_{i,l}(t)\left(d_{\mathrm{KL}}(\hat{\mu}_{i,l}(t), \theta^*_i\kappa^*_l) - f(N_{i,l}(t))\right), \mathcal{G} = \mathcal{H}(t), \mathcal{P}^c(t)\right\}$$

then

$$\sum_{(i,l)\in\mathcal{G}} N_{i,l}(t)\left(d_{\mathrm{KL}}(\hat{\mu}_{i,l}(t),\theta_i^*\kappa_l^*)-f(N_{i,l}(t))\right)$$

$$\geq \inf_{\{\theta_i\}\in\mathcal{T}^c_{\hat{1}(t),\dots,\hat{L}(t)},\{\kappa_l\}\in\mathcal{K}_{\mathrm{all}}:\forall l\in[L]\,d_{\mathrm{KL}}(\hat{\mu}_{\hat{l}(t),l}(t),\theta_{\hat{l}(t)}^*\kappa_l^*)\leq f(N_{\hat{l}(t),l}(t))} \sum_{(i,l)\in[K]\times[L]}\left(d_{\mathrm{KL}}(\hat{\mu}_{i,l}(t),\theta_i\kappa_l)-f(N_{i,l}(t))\right),$$

which implies that at least one of the pairs (i,l) in $\mathcal{G}$ is immediately put into L_N to satisfy the constraints at $(\{\theta_i^*\},\{\kappa_l^*\})$. Therefore, one of the pairs in $\mathcal{G}$ is drawn within K^2 rounds because $|L_N|$ is at most K^2. By using this fact, we have

$$\sum_t \mathbf{1}\left[\mathcal{X}^c(t),\mathcal{W}^c(t),\mathcal{G}=\mathcal{H}(t),\mathcal{P}^c(t),N_{i,l}(t)=n_{i,l}\right]\leq \exp\left(\sum_{(i,l)\in\mathcal{G}} n_{i,l}\left(d_{\mathrm{KL}}(\hat{\mu}_{i,l},\theta_i^*\kappa_l^*)-f(N_{i,l}(t))\right)\right)+K^2.$$

Let $\hat{\mu}_{i,l}^n$ be the empirical estimate of $\mu_{i,l}^*=\theta_i^*\kappa_l^*$ with n draws. Let $s_{i,l}=s_{i,l}(\mu_{i,l}^*)=\sup_{p\in[0,1]}d_{\mathrm{KL}}(p,\mu_{i,l}^*)<\infty$. Letting $\mathbb{P}_{i,l}(x_{i,l})=\mathbb{P}[d_{\mathrm{KL}}(\hat{\mu}_{i,j}^{n_{i,l}},\mu_{i,l}^*)\geq x_{i,l}]$, we have

$$\mathbb{E}\left[\sum_t \mathbf{1}\left[\mathcal{X}^c(t),\mathcal{W}^c(t),\mathcal{G}=\mathcal{H}(t),\bigcap_{(i,l)\in\mathcal{G}} N_{i,l}(t)=n_{i,l}\}\right]\right]$$

$$\leq \int_{\{x_{i,l}\}\in[f(n_{i,l}),s_{i,l}]^{|\mathcal{G}|}}\left(\exp\left(\sum_{(i,l)\in\mathcal{G}} n_{i,l}(x_{i,l}-f(n_{i,l}))\right)+K^2\right)\prod_{(i,l)\in\mathcal{G}}\mathrm{d}(-\mathbb{P}_{i,l}(x_{i,l}))$$

$$= K^2\prod_{(i,l)\in\mathcal{G}} s_{i,l}\mathbb{P}_{i,l}(f(n_{i,l}))+\prod_{(i,l)\in\mathcal{G}}\int_{x_{i,l}\in[f(n_{i,l}),s_{i,l}]} e^{n_{i,l}(x_{i,l}-f(n_{i,l}))}\mathrm{d}(-\mathbb{P}_{i,l}(x_{i,l}))$$

$$= K^2\prod_{(i,l)\in\mathcal{G}} s_{i,l}\mathbb{P}_{i,l}(f(n_{i,l}))$$

$$+\prod_{(i,l)\in\mathcal{G}}\left(\left[-e^{n_{i,l}(x_{i,l}-f(n_{i,l}))}\mathbb{P}_{i,l}(x_{i,l})\right]_{f(n_{i,l})}^{s_{i,l}}+\int_{x_{i,l}\in[f(n_{i,l}),s_{i,l}]} n_{i,l}e^{n_{i,l}(x_{i,l}-f(n_{i,l}))}\mathbb{P}_{i,l}(x_{i,l})\mathrm{d}x_{i,l}\right)$$

(integration by parts)

$$\leq (1+K^2)\prod_{(i,l)\in\mathcal{G}} s_{i,l}\mathbb{P}_{i,l}(f(n_{i,l}))+\prod_{(i,l)\in\mathcal{G}}\int_{x_{i,l}\in[f(n_{i,l}),s_{i,l}]} n_{i,l}e^{n_{i,l}(x_{i,l}-f(n_{i,l}))}e^{-n_{i,l}x_{i,l}}\mathrm{d}x_{i,l}$$

$$\leq (1+K^2)\prod_{(i,l)\in\mathcal{G}} s_{i,l}e^{-n_{i,l}f(n_{i,l})}+\prod_{(i,l)\in\mathcal{G}}\int_{x_{i,l}\in[f(n_{i,l}),s_{i,l}]} n_{i,l}e^{-n_{i,l}f(n_{i,l})}\mathrm{d}x_{i,l}$$

$$= (1+K^2)\prod_{(i,l)\in\mathcal{G}} s_{i,l}e^{-n_{i,l}f(n_{i,l})}+\prod_{(i,l)\in\mathcal{G}} s_{i,l}n_{i,l}e^{-n_{i,l}f(n_{i,l})}. \tag{17}$$

By summing (17) over $\{n_{i,l}\}$,

$$\sum_{t=1}^T \mathbb{P}\left[\mathcal{X}^c(t),\mathcal{W}^c(t),\mathcal{G}=\mathcal{H}(t),\mathcal{P}^c(t)\right]$$

$$\leq \sum_{\{n_{i,l}\}\in\mathbb{N}^{|\mathcal{G}|}}\cdots\sum\left((1+K^2)\prod_{(i,l)\in\mathcal{G}} s_{i,l}e^{-n_{i,l}f(n_{i,l})}+\prod_{(i,l)\in\mathcal{G}} s_{i,l}n_{i,l}e^{-n_{i,l}f(n_{i,l})}\right)$$

$$= O(1).$$

Taking a union bound over $\mathcal{G}\neq\emptyset$ yields

$$\mathbb{P}\left[\mathcal{X}^c(t),\mathcal{W}^c(t),\mathcal{H}(t)\neq\emptyset,\mathcal{P}^c(t)\right]=O(1). \tag{18}$$

We finally obtain

$$\mathbb{P}\left[\mathcal{X}^c(t),\mathcal{W}^c(t)\right]=\mathbb{P}\left[\mathcal{P}(t)\right]+\mathbb{P}\left[\mathcal{X}^c(t),\mathcal{H}(t)=\emptyset,\mathcal{W}^c(t)\right]+\mathbb{P}\left[\mathcal{X}^c(t),\mathcal{W}^c(t),\mathcal{H}(t)\neq\emptyset,\mathcal{P}^c(t)\right]=O(1),$$

where each term is bounded by Lemma 16, Lemma 17, and inequality (18). $\qquad\square$

J Proof of Lemma 11

Following Hogan [16], we define the continuity of a point-to-set map $\Omega : X \to 2^Y$ between metric spaces X and Y as follows: 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 $y^k \to y_0$. 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, Ω is continuous at x_0 if it is closed and open at x_0.

Let $\mathcal{Q}^{\mathrm{L}} = \{\{q_{i,l}\}_{(i,l) \in [K] \times [L]} \in [0, \infty)^{K \times L} : \exists_{\{q_{i,l}\}_{i \in [K], l \in [K] \setminus [L]}}, \{q_{i,l}\}_{(i,l) \in [K] \times [K]} \in \mathcal{Q}\}$ be a restriction of $\mathcal{Q}$ into $K \times L$ dimension. Note that the convexity of $\mathcal{Q}^{\mathrm{L}}$ follows from the convexity of $\mathcal{Q}$. Let a set of feasible solutions restricted to $[K] \times [L]$ space be

$$\mathcal{R}^{\mathrm{L}}_{(1),\dots,(L)}(\{\mu_{i,l}\}, \{\theta_i\}, \{\kappa_l\}, \{\delta_{i,l}\}) = \left\{ \{q_{i,l}\} \in \mathcal{Q}^{\mathrm{L}} : \right.$$

$$\left. \inf_{\{\theta'_i\} \in \mathcal{T}^c_{(1),\dots,(L)}, \{\kappa'_l\} \in \mathcal{K}_{\mathrm{all}} : \forall_{l \in [L]} d_{\mathrm{KL}}(\mu_{(l),l}, \theta'_{(l)} \kappa'_l) \leq \delta_{i,l}} \sum_{(i,l) \in [K] \times [L] : i \neq (l)} q_{i,l} d_{\mathrm{KL}}(\mu_{i,l}, \theta'_i \kappa'_l) \geq 1 \right\}. \quad (19)$$

The set of the optimal solutions $\mathcal{R}^{*,\mathrm{L}}_{(1),\dots,(L)}(\cdot)$ in $K \times L$ dimensions are defined in accordance with $\mathcal{R}^{\mathrm{L}}_{(1),\dots,(L)}(\cdot)$, that is,

$$\mathcal{R}^{*,\mathrm{L}}_{(1),\dots,(L)}(\{\mu_{i,l}\}, \{\theta_i\}, \{\kappa_l\}, \{\delta_{i,l}\}) = \left\{ \{q_{i,l}\} \in \mathcal{Q}^{\mathrm{L}} : \exists_{\{q_{i,l}\}_{i \in [K], l \in [K] \setminus [L]} : \{q_{i,l}\}_{(i,l) \in [K] \times [K]} \in \mathcal{Q}}, \right.$$

$$\left. \sum_{(i,l) \in [K] \times [L]} \Delta_{i,l} q_{i,l} = C^{*,\mathrm{H}}_{(1),\dots,(L)}(\{\mu_{i,l}\}, \{\theta_i\}, \{\kappa_l\}, \{\delta_{i,l}\}) \right\}.$$

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

Lemma 18. (The continuity of the solution function) *The point-to-set map* $\mathcal{R}^{*,\mathrm{L}}_{(1),\dots,(L)}(\{\mu_{i,l}\}, \{\theta_i\}, \{\kappa_l\}, \{\delta_{i,l}\})$ *is continuous at* $(\{\mu_{i,l}\}, \{\theta_i\}, \{\kappa_l\}, \{\delta_{i,l}\}) = (\{\mu^*_{i,l}\}, \{\theta^*_i\}, \{\kappa^*_i\}, \{0\})$.

The continuity and the uniqueness of the optimal solution function $\mathcal{R}^{*,\mathrm{L}}_{(1),\dots,(L)}(\{\mu_{i,l}\}, \{\theta_i\}, \{\kappa_l\}, \{\delta_{i,l}\})$ implies that all solutions of $\mathcal{R}^{*,\mathrm{L}}_{(1),\dots,(L)}(\{\mu_{i,l}\}, \{\theta_i\}, \{\kappa_l\}, \{\delta_{i,l}\})$ approach $\mathcal{R}^{*,\mathrm{L}}_{(1),\dots,(L)}(\{\mu^*_{i,l}\}, \{\theta^*_i\}, \{\kappa^*_i\}, \{0\})$ when $(\{\mu_{i,l}\}, \{\theta_i\}, \{\kappa_l\}, \{\delta_{i,l}\})$ is sufficiently close to $(\{\mu^*_{i,l}\}, \{\theta^*_i\}, \{\kappa^*_i\}, \{0\})$. To prove Lemma 18, we first restate the following three Lemmas of Hogan [16]:

Lemma 19. (Hogan [16, Theorem 10]) *Let g be a real-valued function on $X \times Y$, and $P(x) = \{y \in Y : g(x, y) \leq 0\}$ be a map of feasible solutions. If g is continuous on $x_0 \times Y$, then P is closed at x_0.*

Lemma 20. (Hogan [16, Theorem 13]) *Let $I(x) = \{y \in Y : g(x, y) < 0\} \subset P(x)$. If Y is convex and normed, if g is continuous on $x_0 \times P(x_0)$ and $P(x_0) \subset cl(I(x_0))$, then P is open at x_0, where $cl(\cdot)$ is a closure operator.*

Lemma 21. (Hogan [16, Corollary 8.1]) *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 18. We first show the continuity of the feasible solution function $\mathcal{R}^{\mathrm{L}}_{(1),\dots,(L)}(\cdot)$ at $(\{\mu^*_{i,l}\}, \{\theta^*_i\}, \{\kappa^*_i\}, \{0\})$. Applying Lemma 19 for $P = \mathcal{R}^{\mathrm{L}}_{(1)\dots(L)}$, $x_0 = (\{\mu^*_{i,l}\}, \{\theta^*_i\}, \{\kappa^*_i\}, \{0\})$

23

and

$$g\Big(\big(\{\mu_{i,l}\},\{\theta_i\},\{\kappa_l\},\{\delta_{i,l}\}\big),\{q_{i,l}\}\Big)$$

$$= \sup_{\{\theta_i'\}\in\mathcal{T}^c_{1,\dots,L},\,\{\kappa_l'\}\in\mathcal{K}_{\mathrm{all}}:\forall l\in[L]\,d_{\mathrm{KL}}(\mu_{(l),l},\theta'_{(l)}\kappa'_l)\le\delta_{i,l}}\left(1-\sum_{(i,l)\in[K]\times[L]:i\ne(l)}q_{i,l}d_{\mathrm{KL}}(\mu_{i,l},\theta_i'\kappa_l')\right)$$

yields the closedness of $\mathcal{R}^{\mathrm{L}}_{(1)\dots(L)}$ at $(\{\mu^*_{i,l}\},\{\theta^*_i\},\{\kappa^*_i\},\{0\})$, where the continuity of g follows from the uniform continuity of the KL divergence in the region where $\theta_i'\kappa_l'$ is sufficiently far from $\{0,1\}$. Moreover, let $\{\mu_{i,l}\},\{\theta_i\},\{\kappa_l\},\{\delta_{i,l}\},\{q_{i,l}\}$ be such that $0 = g((\{\mu_{i,l}\},\{\theta_i\},\{\kappa_l\},\{\delta_{i,l}\}),\{q_{i,l}\})$. Then $\{(1-\epsilon)q_{i,l}\}\in\mathcal{Q}^{\mathrm{L}}$ and $g((\{\mu_{i,l}\},\{\theta_i\},\{\kappa_l\},\{\delta_{i,l}\}),\{(1-\epsilon)q_{i,l}\}) = \epsilon$ for any $\epsilon\in[0,1]$, which implies that $\{q_{i,l}\}$ is in $cl(I((\{\mu_{i,l}\},\{\theta_i\},\{\kappa_l\},\{\delta_{i,l}\})))$. By the fact above and the continuity of g, applying Lemma 20 to the same P, x_0, g yields the openness of $\mathcal{R}^{\mathrm{L}}_{(1)\dots(L)}$ at $(\{\mu^*_{i,l}\},\{\theta^*_i\},\{\kappa^*_i\},\{0\})$. The continuity of $\mathcal{R}^{\mathrm{L}}_{(1)\dots(L)}$ follows from its closedness and the openness.

Finally, by using the continuity of $\mathcal{R}^{\mathrm{L}}_{(1)\dots(L)}$ and $C^*_{(1),\dots,(L)}$, and uniform compactness and uniqueness of $\mathcal{R}^{*,\mathrm{L}}_{(1),\dots,(L)}$ at $(\{\mu^*_{i,l}\},\{\theta^*_i\},\{\kappa^*_i\},\{0\})$, applying Lemma 21 to $M = \mathcal{R}^{*,\mathrm{L}}_{(1),\dots,(L)},\Omega = \mathcal{R}^{\mathrm{L}}_{(1)\dots(L)}$, and $f = \mathcal{R}^{*,\mathrm{L}}_{(1),\dots,(L)}$ yields the continuity of $\mathcal{R}^{*,\mathrm{L}}_{(1),\dots,(L)}$ at $(\{\mu^*_{i,l}\},\{\theta^*_i\},\{\kappa^*_i\},\{0\})$. $\qquad\square$

Proof of Lemma 11. By using the continuity of $\mathcal{R}^{*,\mathrm{L}}_{1,\dots,L}(\{\mu^*_{i,l}\},\{\theta^*_i\},\{\kappa^*_i\},\{0\})$ (Lemma 18) and the uniqueness of $\mathcal{R}^{*,\mathrm{L}}_{(1),\dots,(L)}(\{\mu^*_{i,l}\},\{\theta^*_i\},\{\kappa^*_i\},\{0\})$, there exists $\epsilon(\delta)$ such that $\epsilon\to 0$ as $\delta\to +0$ and

$$\sum_{t=1}^{T}\mathbf{1}[\mathcal{Y}_{i,l}(t),\mathcal{Z}_\delta(t)] \le \sum_{n=1}^{T}\mathbf{1}\left[\bigcup_{t=1}^{T}\{\mathcal{Y}_{i,j}(t),\mathcal{Z}_\delta(t),N_{i,l}(t)=n\}\right]$$

$$\le \sum_{n=1}^{T}\mathbf{1}\left[\bigcup_{t=1}^{T}\left\{n/\log t\le(1+\epsilon(\delta))\left(\mathcal{R}^{*,\mathrm{L}}_{(1),\dots,(L)}(\{\mu^*_{i,l}\},\{\theta^*_i\},\{\kappa^*_i\},\{0\})\right)_{i,l}\right\}\right]$$

$$\le (1+\epsilon(\delta))\left(\mathcal{R}^{*,\mathrm{L}}_{(1),\dots,(L)}(\{\mu^*_{i,l}\},\{\theta^*_i\},\{\kappa^*_i\},\{0\})\right)_{i,l}\log T + 1,$$

where $\left(\mathcal{R}^{*,\mathrm{L}}_{(1),\dots,(L)}(\{\mu^*_{i,l}\},\{\theta^*_i\},\{\kappa^*_i\},\{0\})\right)_{i,l}$ is the corresponding (i,l) entry, which completes the proof. $\qquad\square$

K   Proof of Lemma 12

Lemma 22. *Let $C_2 > 0$ be arbitrary. Assume that $|\hat\theta_i(t)-\hat\theta_j(t)|\ge C_2$ for any $i,j\in[K]\times[K]$ and $|\hat\kappa_l(t)-\hat\kappa_m(t)|\ge C_2$ for any $l\ne m\in[L]$ hold. Moreover, assume that $f(N_{i,l}(t))\le C_2^2/4$ holds for any $(i,l)\in[K]\times[L]$. Then, for any optimal solution, the following inequality holds for any $(i,l)\notin\{(\hat1(t),1),(\hat2(t),2),\dots,(\hat L(t),L)\}$:*

$$(\mathcal{R}^{*,\mathrm{L}}_{\hat1(t),\dots,\hat L(t)})_{i,l}(\{\hat\mu_{i,l}(t)\},\{\hat\theta_i(t)\},\{\hat\kappa_i(t)\},\{f(N_{i,l}(t))\}) \le \frac{4LK}{C_2^4}.$$

Proof of Lemma 22. If $\{\theta_i'\}\in\mathcal{T}^c_{\hat1(t),\dots,\hat L(t)}$, then there exists i',j' such that $(\theta'_{i'}-\theta'_{j'})(\hat\theta_i(t)-\hat\theta_j(t))\le 0$, which implies $\max(|\theta'_{i'}-\hat\theta_{i'}(t)|,|\theta'_{j'}-\hat\theta_{j'}(t)|)\ge C_2/2$. Without loss of generality let

24

$|\theta'_{i'} - \hat{\theta}_{i'}(t)| \geq C_2/2$. Then, for any $\{\theta'_i\}, \{\kappa'_l\}$,

$$\sum_{i,l} q_{i,l}\left(d_{\mathrm{KL}}(\hat{\mu}_{i,l}(t), \theta'_i \kappa'_l) - f(N_{i,l}(t))\right)_+$$

$$\geq q_{i',1}\left(d_{\mathrm{KL}}(\hat{\theta}_{i'}(t), \theta'_{i'}) - f(N_{i',1}(t))\right)_+$$

$$\geq q_{i',1}\left(2(\hat{\theta}_{i'}(t) - \theta'_{i'})^2 - f(N_{i',1}(t))\right)_+$$

(by Pinsker's inequality)

$$\geq q_{i',1}\left(C_2^2/2 - C_2^2/4\right)_+$$

$$= q_{i',1}C_2^2/4,$$

which implies that

$$\left\{q_{i,l} = \frac{4}{C_2^2}\right\}_{(i,l)\in[K]\times[L]} \in \mathcal{R}^{\mathrm{L}}_{\hat{1}(t),\ldots,\hat{L}(t)}(\{\hat{\mu}_{i,l}(t)\}, \{\hat{\theta}_i(t)\}, \{\hat{\kappa}_i(t)\}, \{f(N_{i,l}(t))\}).$$

The estimated regret on the basis of $\{\hat{\theta}_i(t)\}, \{\hat{\kappa}_l(t)\}$ increases at least $\min_{i\neq j}\min_{l\neq m}|\hat{\theta}_i(t) - \hat{\theta}_j(t)||\hat{\kappa}_l(t) - \hat{\kappa}_m(t)| \geq C_2^2$ when we draw arms that are not $\hat{1}(t), \ldots \hat{L}(t)$. Therefore, $q_{i,l}$ for $i \neq \hat{l}(t)$ is bounded by $\frac{4LK}{C_2^4}$. $\qquad\square$

Let

$$\mathcal{A}(t) = \bigcap_{i,l}\left\{d_{\mathrm{KL}}(\hat{\mu}_{i,l}(t), \theta^*_i \kappa^*_l) \leq f(N_{i,l}(t))\right\}.$$

Then, the following lemma holds.

Lemma 23. *For sufficiently small $\delta > 0$, we have*

$$\sum_{t=1}^{T}\mathbb{P}\left[N_{i,l}(t) > \alpha\sqrt{\log t}, \mathcal{Z}^c_\delta(t), \mathcal{A}(t)\right] = O(1).$$

Proof of Lemma 23. Let

$$\mathcal{B}_\delta(t) = \bigcap_{i,l}\left\{f(N_{i,l}(t)) < \delta^2\right\}.$$

Then,

$$\sum_{t=1}^{T}\mathbb{P}\left[\mathcal{X}^c(t), \mathcal{B}^c_\delta(t)\right]$$

$$\leq \sum_{(i,l)\in[K]\times[L]}\sum_{t=1}^{T}\mathbb{P}\left[N_{i,j}(t) \geq \alpha\sqrt{\log t}, f(N_{i,l}(t)) > \delta^2\right]$$

$$\leq \sum_{(i,l)\in[K]\times[L]}\sum_{t=1}^{T}\mathbb{P}\left[\gamma\frac{1}{\alpha^{1/2}(\log t)^{1/4}} > \delta^2\right]$$

$$\leq K^2 e^{\left(\frac{\gamma^4}{\alpha^2\delta^8}\right)} = O(1). \tag{20}$$

Moreover, $\mathcal{A}(t)$ implies

$$\bigcap_{i,l}\left\{d_{\mathrm{KL}}(\hat{\mu}_{i,l}(t), \hat{\theta}_i(t)\hat{\kappa}_l(t)) \leq f(N_{i,l}(t))\right\},$$

and by using the Pinsker's inequality we have

$$|\hat{\mu}_{i,l}(t) - \hat{\theta}_i(t)\hat{\kappa}_l(t)| \leq \frac{1}{2}\sqrt{f(N_{i,l}(t))}$$

$$|\hat{\mu}_{i,l}(t) - \theta^*_i \kappa^*_l| \leq \frac{1}{2}\sqrt{f(N_{i,l}(t))},$$

25

for any $(i, l) \in [K] \times [L]$, which implies

$$|\hat{\theta}_i(t)\hat{\kappa}_l(t) - \theta_i^* \kappa_l^*| \leq \sqrt{f(N_{i,l}(t))}$$

for any $(i, l) \in [K] \times [L]$. Therefore,

$$\sum_{t=1}^{T} \mathbb{P}\left[\mathcal{X}^c(t), \mathcal{Z}_\delta^c(t), \mathcal{A}(t)\right]$$

$$\leq \sum_{t=1}^{T} \mathbb{P}\left[\mathcal{X}^c(t), \mathcal{B}_\delta^c(t)\right] + \sum_{t=1}^{T} \mathbb{P}\left[\mathcal{X}^c(t), \mathcal{B}_\delta(t), \mathcal{Z}_\delta^c(t), \mathcal{A}(t)\right]$$

$$\leq O(1) + \mathbb{P}\left[\mathcal{Z}_\delta^c(t), \bigcap_{(i,l)\in[K]\times[L]} \{f(N_{i,l}(t)) \leq \delta^2, |\hat{\theta}_i(t)\hat{\kappa}_l(t) - \theta_i^* \kappa_l^*| \leq \sqrt{f(N_{i,l}(t))}\}\right]$$

(by inequality (20))

$$\leq O(1) + \mathbb{P}\left[\mathcal{Z}_\delta^c(t), \bigcap_{(i,l)\in[K]\times[L]} \{|\hat{\theta}_i(t)\hat{\kappa}_l(t) - \theta_i^* \kappa_l^*| \leq \delta\}\right] = O(1) + 0.$$

$$\square$$

Lemma 24. *The following inequality holds:*

$$\sum_{t=1}^{T} \mathbf{1}\left[\mathcal{X}^c(t), \mathcal{B}_{(\beta/(\log\log t))/4}^c(t)\right] = O(1).$$

Proof of Lemma 24. We have

$$\sum_{(i,l)\in[K]\times[L]} \sum_{t=1}^{T} \mathbf{1}\left[\mathcal{X}^c(t), \mathcal{B}_{(\beta/(\log\log t))/4}^c(t)\right] \leq \sum_{(i,l)\in[K]\times[L]} \sum_{t=1}^{T} \mathbf{1}\left[\frac{\gamma}{\sqrt{\alpha}\sqrt{\log t}} > \frac{\beta^2}{4(\log\log t)}\right]$$

$$\leq \sum_{(i,l)\in[K]\times[L]} \sum_{t=1}^{T} \mathbf{1}\left[\frac{\log t}{(\log\log t)^4} < \frac{2^8 \gamma^4}{\alpha^2 \beta^4}\right] = O(1).$$

$$\left(\text{by } \frac{\log t}{(\log\log t)^8} \to \infty \text{ as } t \to \infty\right)$$

$$\square$$

26

Proof of Lemma 12. We have

$$\sum_{t=1}^{T} \mathbb{P}\left[\mathcal{X}^c(t), \mathcal{Y}_{i,l}(t), \mathcal{Z}_\delta^c(t)\right]$$

$$\leq \sum_{t=1}^{T} \mathbb{P}\left[\mathcal{X}^c(t) \bigcap_{(i',l')\in[K]\times[L]} N_{i',l'}(t) \geq (\log\log T)^{1/3}, \mathcal{Y}_{i,l}(t), \mathcal{Z}_\delta^c(t)\right]$$

$$+ \sum_{(i',l')\in[K]\times[L]} \sum_{t=1}^{T} \mathbb{P}[N_{i',l'}(t) \leq (\log\log T)^{1/3}, N_{i',l'}(t) \geq \alpha\sqrt{\log t}]$$

$$\leq \sum_{t=1}^{T} \mathbb{P}\left[\mathcal{X}^c(t), \bigcap_{(i',l')\in[K]\times[L]} N_{i',l'}(t) \geq (\log\log T)^{1/3}, \mathcal{Y}_{i,l}(t), \mathcal{A}^c(t)\right]$$

$$+ \sum_{(i',l')\in[K]\times[L]} \sum_{t=1}^{T} \mathbb{P}[N_{i',l'}(t) \leq (\log\log T)^{1/3}, N_{i',l'}(t) \geq \alpha\sqrt{\log t}]$$

$$+ \sum_{t=1}^{T} \mathbb{P}\left[\mathcal{X}^c(t), \mathcal{Z}_\delta^c(t), \mathcal{A}(t)\right]. \tag{21}$$

Here, the second term of (21) is bounded as:

$$\sum_{(i',l')\in[K]\times[L]} \sum_{t=1}^{T} \mathbb{P}[N_{i',l'}(t) \leq (\log\log T)^{1/3}, N_{i',l'}(t) \geq \alpha\sqrt{\log t}] \leq K^2 e^{\alpha^{-2}(\log\log T)^{2/3}} = o(\log T). \tag{22}$$

The third term of (21) is $O(1)$ by Lemma 23. In the following we bound the first term of (21).

$$\sum_{t=1}^{T} \mathbb{P}\left[\mathcal{X}^c(t), \bigcap_{(i',l')\in[K]\times[L]} N_{i',l'}(t) \geq (\log\log T)^{1/3}, \mathcal{Y}_{i,l}(t), \mathcal{A}^c(t)\right]$$

$$\leq \sum_{t=1}^{T} \mathbb{P}\left[\mathcal{X}^c(t), \bigcap_{(i',l')\in[K]\times[L]} N_{i',l'}(t) \geq (\log\log T)^{1/3}, \mathcal{Y}_{i,l}(t), \mathcal{A}^c(t), \mathcal{B}_{(\beta/(\log\log t))/4}(t)\right] + O(1)$$

(by Lemma 24)

$$\leq \sum_{n=1}^{T} \mathbb{P}\left[\bigcup_{t=n}^{T}\left\{\mathcal{X}^c(t), \bigcap_{(i',l')\in[K]\times[L]} N_{i',l'}(t) \geq (\log\log T)^{1/3}, \mathcal{Y}_{i,l}(t), \mathcal{A}^c(t), \mathcal{B}_{(\beta/(\log\log t))/4}(t), N_{i,l}(t) = n\right\}\right]$$

$$+ O(1)$$

$$\leq \sum_{n=1}^{\log T (4LK(\log\log T/\beta)^4/2)} \mathbb{P}\left[\bigcup_{t=1}^{T}\left\{N_{i',l'}(t) \geq (\log\log T)^{1/3}, \mathcal{A}^c(t), N_{i,l}(t) = n\right\}\right] + O(1)$$

(by the fact that $\mathcal{X}^c(t)$ combined with Lemma 22 with $C_2 = \beta/\log\log t$

$$\text{imply that } q_{i,l} \leq 4LK(\log\log T/\beta)^4 \text{ for any } (l) \neq i)$$

$$\leq e^{-\Omega((\log\log T)^{1/3})} O((\log T)(\log\log T)^4) = o(\log T), \tag{23}$$

27

where, in the last line we used the fact that

$$\mathbb{P}\left[\bigcup_{t=1}^{T}\left\{\mathcal{X}^c(t), \mathcal{A}^c(t), \bigcap_{(i',l')\in[K]\times[L]} N_{i',l'}(t) \geq (\log\log T)^{1/3}\right\}\right]$$

$$\leq \sum_{(i',l')\in[K]\times[L]} \sum_{n=(\log\log T)^{1/3}}^{T} \left(\mathbb{P}[|\hat{\mu}_{i',l'}^n - \mu_{i',l'}^*| > \delta] + \mathbf{1}\{\delta \geq f(\alpha\sqrt{\log t})\}\right)$$

$$\leq \sum_{(i',l')\in[K]\times[L]} \sum_{n=(\log\log T)^{1/3}}^{T} \mathbb{P}[|\hat{\mu}_{i',l'}^n - \mu_{i',l'}^*| > \delta] + O(1)$$

$$\leq \sum_{(i',l')\in[K]\times[L]} \sum_{n=(\log\log T)^{1/3}}^{T} 2e^{-2n\delta} = e^{-\Omega((\log\log T)^{1/3})} + O(1).$$

Combining (21), (22), and (23) completes the proof. □

L   Notation

Table 2 summarizes the notation appeared in this paper.

Table 2: List of notation that are used in more than one sections in this paper.

T	Number of rounds (Sec. 2)
K	Number of arms (Sec. 2)
L	Number of slots (Sec. 2)
θ_i^*	Parameter associated with arm $i \in [K]$ (Sec. 2)
κ_l^*	Parameter associated with slot $l \in [L]$ (Sec. 2)
$\mu_{i,l}^*$	$= \theta_i^* \kappa_l^*$ (Sec. 2)
$\Delta_{i,l}$	$= \theta_l^* \kappa_l^* - \theta_i^* \kappa_l^*$ (Sec. 2)
$I(t) = (I_1(t), \ldots, I_L(t))$	List of arms selected at round t (Sec. 2)
$N_{i,l}(t)$	Number of rounds before t at which arm i was in slot l (Sec. 2)
$\hat{\mu}_{i,l}(t)$	Empirical mean of pair (i,l) at round t (Sec. 2)
$(1), \ldots, (K)$	A permutation of $1, \ldots, K$ (Sec. 3)
$\mathcal{T}_{\mathrm{all}}$	Set of possible values of parameters on the arms (Sec. 3)
$\mathcal{K}_{\mathrm{all}}$	Set of possible values of parameters on the slots (Sec. 3)
$\mathcal{T}_{(1),\ldots,(L)}$	Subset of $\mathcal{T}_{\mathrm{all}}$ such that the i-th best arm is (i) (Sec. 3)
$\mathcal{T}_{(1),\ldots,(L)}^c$	$\mathcal{T}_{\mathrm{all}} \setminus \mathcal{T}_{(1),\ldots,(L)}$ (Sec. 3)
$d_{\mathrm{KL}}(p, q)$	$p \log(p/q) + (1-p) \log((1-p)/(1-q))$ (Sec. 3)
$\mathcal{Q}$	Subspace of $[0,\infty)^{[K] \times [K]}$ that corresponds to the number of draws (Sec. 3)
$\mathcal{R}_{(1),\ldots,(L)}(\{\mu_{i,l}\}, \{\theta_i\}, \{\kappa_l\})$	Feasible solutions under parameters $\{\mu_{i,l}\}, \{\theta_i\}, \{\kappa_l\}$ and top-L arms $(1), \ldots, (L)$ (Sec. 3)
$C_{(1),\ldots,(L)}^*(\{\mu_{i,l}\}, \{\theta_i\}, \{\kappa_l\})$	Coefficient of optimal regret bound (Sec. 3)
$\mathcal{R}_{(1),\ldots,(L)}^*(\{\mu_{i,l}\}, \{\theta_i\}, \{\kappa_l\})$	Optimal solutions (Sec. 3)
$\alpha > 0$	Parameter of PMED algorithm (Sec. 4)
L_C	List of L-allocations that are drawn at current loop of PMED (Sec. 4)
L_N	List of L-allocations that are drawn at next loop of PMED (Sec. 4)
v_m^{mod}	An L-allocation (Sec. 4)
$\{\hat{\theta}_i(t)\}_{i=1}^K, \{\hat{\kappa}_l(t)\}_{l=1}^L$	MLE. Note that $\hat{\mu}_{i,l}(t) \neq \hat{\theta}_i(t)\hat{\kappa}_l(t)$. (Sec. 4)
$\tilde{N}_{i,l}$	$K \times K$ variables that correspond to the estimated amount of exploration on pair (i,l) (Sec. 4)
e_v	A permutation matrix (Sec. 4)
$e_{v,i,l}$	(i,l) entry of e_v (Sec. 4)
$(\hat{1}(t), \ldots, \hat{L}(t))$	Top-L arms estimated from MLE (Sec. 4)
$(x)_+$	$= \max(x, 0)$ (Sec. 7)
$\beta, \gamma > 0$	Parameters of PMED-Hinge algorithm (Sec. 7)
$f(n)$	$= \gamma/\sqrt{n}$ (Sec. 7)
$\mathcal{R}_{(1),\ldots,(L)}^{\mathrm{H}}(\{\mu_{i,l}\}, \{\theta_i\}, \{\kappa_l\}, \{\delta_{i,l}\})$	Feasible solutions of the modified objective (Sec. 7)
$C_{(1),\ldots,(L)}^{*,\mathrm{H}}(\{\mu_{i,l}\}, \{\theta_i\}, \{\kappa_l\}, \{\delta_{i,l}\})$	Coefficient of optimal regret bound of modified objective (Sec. 7)
$\mathcal{R}_{(1),\ldots,(L)}^{*,\mathrm{H}}(\{\mu_{i,l}\}, \{\theta_i\}, \{\kappa_l\}, \{\delta_{i,l}\})$	Optimal solutions of the modified objective (Sec. 7)
$\mathcal{W}(t)$	Event that top-L arms is misidentified (Sec. G)
$\mathcal{J}_{i,l}(t)$	Event that pair (i,l) is put into L_N (Sec. G)
$\mathcal{X}(t)$	Event that at least one arm is put into L_N before Line 18 (Sec. G)
$\mathcal{Y}_{i,l}(t)$	Eevent that pair (i,l) is put into L_N in Line 18 (Sec. G)
$\mathcal{Z}_\delta(t)$	Event that estimator $\{\hat{\theta}_i(t), \hat{\kappa}_l(t)\}$ is δ-close to the set of true values (Sec. G)
$R_{i,l}^*$	(i,l) entry of the optimal solution $\mathcal{R}_{1,\ldots,L}^*(\{\mu_{i,l}^*\}, \{\theta_i^*\}, \{\kappa_l^*\})$ (Sec. G)
$\hat{\mu}_{i,l}^n$	$\hat{\mu}_{i,l}(t)$ when $N_{i,l}(t) = n$ (Sec. G)
$\mathcal{C}(t)$	Event that hinge of MLE is zero (Sec. H)
$\mathcal{H}(t)$	Subset of entries $(i,l) \in [K] \times [L]$ such that hinge of true parameters is zero (Sec. I)