

# Continuous K-Max Bandits

Yu Chen<sup>1\*</sup> Siwei Wang<sup>2\*</sup> Longbo Huang<sup>1</sup> Wei Chen<sup>2</sup>✉

<sup>1</sup>IIIS, Tsinghua University

<sup>2</sup>Microsoft Research Asia

chenyu23@mails.tsinghua.edu.cn

siweiwang@microsoft.com

longbohuang@tsinghua.edu.cn

weic@microsoft.com

## Abstract

We study the  $K$ -Max combinatorial multi-armed bandits problem with continuous outcome distributions and weak value-index feedback: each base arm has an unknown continuous outcome distribution, and in each round the learning agent selects  $K$  arms, obtains the maximum value sampled from these  $K$  arms as reward and observes this reward together with the corresponding arm index as feedback. This setting captures critical applications in recommendation systems, distributed computing, server scheduling, etc. The continuous  $K$ -Max bandits introduce unique challenges, including discretization error from continuous-to-discrete conversion, non-deterministic tie-breaking under limited feedback, and biased estimation due to partial observability. Our key contribution is the computationally efficient algorithm DCK-UCB, which combines adaptive discretization with bias-corrected confidence bounds to tackle these challenges. For general continuous distributions, we prove that DCK-UCB achieves a  $\tilde{\mathcal{O}}(T^{3/4})$  regret upper bound, establishing the first sublinear regret guarantee for this setting. Furthermore, we identify an important special case with exponential distributions under full-bandit feedback. In this case, our proposed algorithm MLE-Exp enables  $\tilde{\mathcal{O}}(\sqrt{T})$  regret upper bound through maximal log-likelihood estimation, achieving near-minimax optimality.

## 1 Introduction

Multi-armed bandits (MABs) provide a powerful framework for sequential decision-making under uncertainty, balancing exploration and exploitation to maximize cumulative rewards. Among its variants, Combinatorial MABs (CMABs) Cesa-Bianchi and Lugosi (2012); Chen et al. (2013) have gained significant attention due to applications in online advertising, networking, and influence maximization (Gai et al., 2012; Kveton et al., 2015a; Chen et al., 2009, 2013). In CMABs, an agent selects a subset of arms as the combinatorial action for each round, and the environment will return the reward signal according to the outcome of selected arms.

As a popular variant, *K*-Max Bandits (Goel et al., 2006; Gopalan et al., 2014) focuses on the maximum outcomes within a selected subset of  $K$  arms. This framework naturally captures real-world scenarios where the decision quality only depends on the extreme singular outcomes. For example, in *recommendation system*, modern ad platforms must select  $K$  products to display

---

\* denotes equal contributions. Corresponding author: Wei Chen (weic@microsoft.com)

from a pool of candidates, where the customer will select the most preferred one. In this case, the extreme preference reflects the recommendation efficiency of selection. Likewise, in *distributed computing tasks*, a scheduler may choose  $K$  servers for parallel processing, and the overall completion metric depends on the server with the fastest response while feedback from others can be overshadowed.

Motivated by the prevalence of continuous real-valued signals (e.g., ratings of selected products in online advertising, job completion time in distributed computing, and latency in server scheduling) and the fact that only partial observations may be accessible in modern applications, we study the *Continuous  $K$ -Max Bandits* problem with *value-index feedback*. Here, each arm has an unknown continuous distribution, and upon selecting a set of  $K$  arms, the learner only observes the maximum outcome among the chosen arms alongside the index of the arm that attained that maximum value.

While  $K$ -Max bandits have received considerable attention (Simchowitz et al., 2016; Chen et al., 2016a; Agarwal et al., 2020), existing theoretical works face three critical limitations when tackling the continuous distribution and value-index feedback: First, most existing algorithms require semi-bandit feedback (Chen et al., 2016a; Simchowitz et al., 2016; Wang and Chen, 2017), while practical systems often restrict observations to the winning arm’s index and value. The key challenge here is that the observation under semi-bandit feedback is unbiased, while the observation under value-index feedback is biased, since we only observe an outcome when it is the winner. Second, greedy approaches based on submodular optimization (Streeter and Golovin, 2008; Fourati et al., 2024) face inherent  $(1 - 1/e)$  approximation limits (Nemhauser et al., 1978), which results in weaker regret guarantees. Third, existing solutions to  $K$ -Max bandits mostly assume binary (Simchowitz et al., 2016) or finitely supported outcomes (Wang et al., 2023). The continuous nature of real-world outcomes, along with the value-index feedback, introduces challenges in discretization error and learning efficiency tradeoff, biased estimators due to nondeterministic tie-breaking under discretization, etc.

Our primary contribution is a novel framework for Continuous  $K$ -Max Bandits with value-index feedback that addresses these challenges through two key technical innovations: **(i)** We formalize the discretization of continuous  $K$ -Max bandits into a discrete  $K$ -Max bandits (see Sections 4.1 to 4.3). In this process, we control the error term from discretization and establish the utilization of the efficient offline  $\alpha$ -approximated optimization oracle for any  $\alpha < 1$ , avoiding the unacceptable approximation error by traditional greedy algorithms that leads to a linear regret. **(ii)** Due to the nondeterministic tie-breaking effect arising from the continuous-to-discrete transformation under value-index feedback, the agent cannot achieve an unbiased estimation under computationally tractable discretization (as detailed in Section 4.4). We develop a bias-corrected discretization method and a novel concentration analysis with bias-aware error control (Lemma 4.5) that jointly manage estimation variance and discretization-induced bias to achieve sublinear regret.

**Contributions.** Our novel techniques lead to the development of the DCK-UCB algorithm (Algorithm 1) for  $K$ -Max bandits with general continuous distributions. The DCK-UCB algorithm achieves a regret upper bound of  $\tilde{\mathcal{O}}(T^{3/4})$ , and is efficient both computationally and statistically. To the best of our knowledge, this represents the first computationally tractable algorithm with a sublinear guarantee for continuous  $K$ -Max bandits under value-index feedback.

We further consider exponential  $K$ -Min bandits, a special case of continuous  $K$ -Max bandits, where outcomes of every arm follow exponential distributions. By utilizing the special property of exponential distributions, we avoid using a discretization method and biased estimators. Following this idea, we adapt the maximum log-likelihood estimation algorithm MLE-Exp (Algorithm 2), and propose an algorithm that achieves a better  $\tilde{\mathcal{O}}(\sqrt{T})$  regret upper bound, which is

nearly minimax optimal.

**Paper organization.** Section 2 introduces related works of continuous  $K$ -Max bandits. Section 3 formalizes the continuous  $K$ -Max bandits problem. Section 4 describes our algorithm and analysis for general continuous  $K$ -Max bandits. Section 5 details the exponential distribution special case and an MLE-based algorithm with better regret guarantees. Complete proofs of Sections 4 and 5 are presented in the appendices.

**Notations.** For any integer  $n \geq 1$ , we use  $[n]$  to denote the set  $\{1, 2, \dots, n\}$ . The notation  $\mathcal{O}$  is used to suppress all constant factors, while  $\tilde{\mathcal{O}}$  is used to further suppress all logarithmic factors. Bold letters such as  $\mathbf{x}$  are typically used to represent a set of elements  $\{x_i\}$ . Unless expressly stated,  $\log(x)$  refers to the natural logarithm of  $x$ . Throughout the text,  $\{\mathcal{F}_t\}_{t=0}^T$  is used to denote the natural filtration; that is,  $\mathcal{F}_t$  represents the  $\sigma$ -algebra generated by all random observations made within the first  $t$  time slots.

## 2 Related Works

The  $K$ -Max bandit problem represents a significant departure from traditional Combinatorial Multi-Armed Bandits (CMABs) (Cesa-Bianchi and Lugosi, 2012; Chen et al., 2013). While standard CMAB frameworks only need to learn the expected outcomes of every arms (Chen et al., 2013, 2014; Kveton et al., 2015b; Combes et al., 2015; Liu et al., 2023b),  $K$ -Max bandits, whose reward signal is the maximum outcome within selected arms (Goel et al., 2006; Gopalan et al., 2014), require to learn more information about the probability distribution of every arms, and necessitate novel approaches to balance the exploration-exploitation tradeoff. Existing literature on  $K$ -Max bandits can be categorized by feedback type as follows.

**Value Feedback** Certain scenarios involve the environment returning only the numerical reward, which corresponds to the maximum outcome of selected arms, known as the *full-bandit feedback*. Gopalan et al. (2014) obtained the regret upper bounds  $\mathcal{O}\left(\sqrt{\binom{N}{K}T}\right)$  for the  $K$ -Max bandits through a Thompson Sampling scheme. However, their approach requires the ground truth parameter to be in a known finite set, and their regret scales exponentially with  $K$ . Simchowitz et al. (2016) considered the pure exploration task while their results are limited to Bernoulli outcome distributions. Streeter and Golovin (2008); Yue and Guestrin (2011); Nie et al. (2022); Fourati et al. (2024) investigated the submodular maximization perspective and yield  $(1 - 1/e)$ -approximation regret guarantees via greedy selection, and  $K$ -Max bandits naturally satisfy the submodular assumption. However, these approximation regret guarantees lead to linear regret when the baseline policy is the true optimal subset (Eq. (1) in this paper). It still remains open on achieving sublinear regret bounds in general  $K$ -Max bandits with full-bandit feedback.

**Value-Index Feedback** In this case, the feedback provides both the maximum outcome (reward) and the corresponding arm index. It looks similar to a CMAB problem with probabilistic triggering feedback (Wang and Chen, 2017; Liu et al., 2023b, 2024b), i.e., with certain probability, we observe arm  $i$  to be the winner, and also get an observation on arm  $i$ 's outcome. The main difference is that in these CMAB researches, it is often assumed that conditioning on we observe arm  $i$ 's outcome, the random distribution of this observed outcome is the same as the real outcome of arm  $i$  without such condition (at least the mean should be the same). However,

this is not the case in  $K$ -Max bandits, i.e., conditioning on arm  $i$  being the winner, the observed outcome of arm  $i$  must have some bias to the real outcome. Under this feedback protocol, the most related work of (Wang et al., 2023) considered the discrete  $K$ -Max bandits with a finite outcome support and a deterministic tie-breaking rule, achieving an  $\tilde{\mathcal{O}}(\sqrt{T})$  regret upper bound. Simchowitz et al. (2016) also investigated the value-index feedback for Bernoulli outcomes. However, their algorithm cannot work for the continuous case studied in this paper due to non-zero discretization error and the nondeterministic tie-breaking.

**Semi-Bandit Feedback** Semi-bandit feedback reveals the outcome of *every* selected arm in the subset, providing the learner with detailed observations to estimate each arm's distribution. Several studies (Simchowitz et al., 2016; Jun et al., 2016; Chen et al., 2016a,b; Slivkins et al., 2019) have leveraged this rich feedback to achieve  $\mathcal{O}(\sqrt{T})$  regret upper bounds for discrete or continuous  $K$ -Max bandits with unbiased estimation for every arm's outcome distribution. However, due to the abundant unbiased observation, their proposed algorithm becomes very different from ours and cannot be applied to our setting.

### 3 Preliminaries

We study the *continuous  $K$ -Max bandits*, denoted as  $\mathcal{B}^*$ , where an agent interacts with  $N$  arms  $\mathcal{A} = [N]$ . For each arm  $i \in [N]$ , there is a corresponding continuous random distribution  $D_i$  such that  $X_i \sim D_i$ , where  $X_i$  is the outcome of arm  $i$ .

The agent will play  $T$  rounds in total. At each time step  $t \in [T]$ , the agent needs to select an action  $S_t$  from the feasible action set  $\mathcal{S} = \{S \subseteq \mathcal{A} \mid |S| = K\}$ , i.e., a subset of  $\mathcal{A}$  with size  $K$ . Here  $1 < K < N$  is a given constant. After selecting  $S_t$ , the environment first samples outcomes  $X_i(t) \sim D_i$  for all  $i \in S_t$ , and all the random variables  $X_i(t)$  (for different  $i, t$  pairs) are sampled independently. Then the environment returns *value-index feedback*  $(r_t, i_t)$ , where  $r_t = \max_{i \in S_t} X_i(t)$  is the maximum outcome, and  $i_t = \operatorname{argmax}_{i \in S_t} X_i(t)$  is the index of the arm that achieves this maximum outcome. Besides,  $r_t$  is also the reward of the agent in time step  $t$ . Note that ties, shall they occur (we denote this event as  $\neg\mathcal{E}_0$ ), are resolved in an arbitrary manner, although  $\mathbb{P}[\neg\mathcal{E}_0] = 0$  since  $D_i$ 's are continuous distributions. We denote the expected reward of an action  $S$  as  $r^*(S)$ , which is given by:

$$r^*(S) := \mathbb{E}[\max\{X_i : i \in S\}] = \int_0^1 r \cdot d\mathbb{P}_{\max\{X_i : i \in S\}}(r).$$

The objective of the  $K$ -Max bandits is to select actions  $S_t$  properly to maximize the cumulative reward  $\sum_{t=1}^T r^*(S_t)$  in  $T$  rounds.

Let  $S^*$  denote the optimal action  $S^* := \operatorname{argmax}_{S \in \mathcal{S}} r^*(S)$ . We evaluate the performance of the agent by the regret metric, which is defined by

$$\mathcal{R}(T) := \mathbb{E} \left[ \sum_{t=1}^T r^*(S^*) - r^*(S_t) \right], \quad (1)$$

where the expectation is taken over the uncertainty of  $\{S_t\}_{t=1}^T$ .

## 4 Algorithm for $K$ -Max Bandits with General Continuous Distribution

We now present our solution framework for continuous  $K$ -Max bandits, beginning with the fundamental regularity condition that enables discretization-based learning:

**Assumption 4.1.** Each outcome distribution  $D_i$  is supported on  $[0, 1]$  with a bi-Lipschitz continuous cumulative distribution function (CDF)  $F_i$ . Specifically, there exists  $L \geq 1$  such that for any  $i \in [N]$  and  $0 \leq v < u \leq 1$ :

$$\frac{1}{L}(u - v) \leq F_i(u) - F_i(v) \leq L(u - v).$$

Many studies on MAB or CMAB consider  $[0, 1]$ -supported arms (Abbasi-Yadkori et al., 2011; Chen et al., 2013; Slivkins et al., 2019; Lattimore and Szepesvári, 2020). The bi-Lipschitz continuity is also common in practice (Li et al., 2017; Wang et al., 2019; Liu et al., 2023a) and satisfied by many distributions such as (truncated) Gaussians, mixed uniforms, Beta distributions, etc.

### 4.1 The Discretization of Countinuous $K$ -Max Bandits

Since it is complex to estimate the general continuous distributions, a natural idea is to perform *discretization* with granularity  $\epsilon$ . Below, we define the *discrete  $K$ -Max bandits* (called  $\bar{\mathcal{B}}$ ) converted from the continuous  $K$ -Max bandits  $\mathcal{B}^*$ , where each discrete arm's outcome  $\bar{X}_i$  is discretized from the continuous random variable  $X_i$  under  $\epsilon$ :

$$\bar{X}_i = \sum_{j \in [M]} \mathbb{1}[X_i \in M_j] \cdot v_j, \quad (2)$$

where  $M = \lceil 1/\epsilon \rceil^1$  is the number of discretization bins,  $M_j := [(j-1)\epsilon, j\epsilon)$  is the  $j$ -th bin, and  $v_j := (j-1)\epsilon$  is the approximate value of  $j$ -th bin. We also let  $M_{\leq j} = \cup_{j' \leq j} M_{j'}$  and  $M_{\geq j} = \cup_{j' \geq j} M_{j'}$ . For simplicity, we denote  $p_{i,j}^*$  as the probability that  $X_i$  falls in  $M_j$ . For every  $i \in [N]$  and  $j \in [M]$ ,

$$p_{i,j}^* := \mathbb{P}[X_i \in M_j] = \mathbb{P}[\bar{X}_i = v_j].$$

Therefore,  $\bar{\mathcal{B}}$  only depends on the discrete probability set  $\mathbf{p}^* = \{p_{i,j}^* : i \in [N], j \in [M]\}$ . Moreover, we set  $\bar{r}(S; \mathbf{p}^*)$  as the expected reward of an action  $S$  in discrete  $K$ -Max bandits under the probability set  $\mathbf{p}^*$ :

$$\bar{r}(S; \mathbf{p}^*) = \sum_{j \in [M]} v_j \cdot \mathbb{P}\left[\max_{i \in S} (\bar{X}_i) = v_j\right]$$

A key observation is that  $\max_{i \in S} (\bar{X}_i) = v_j$  is equivalent to  $\max_{i \in S} (X_i) \in M_j$ . This means  $\mathbb{P}[\max_{i \in S} (\bar{X}_i) = v_j] = \mathbb{P}[\max_{i \in S} (X_i) \in M_j]$ , which gives an upper bound for the discretization error as follows. The formal version is provided by Lemma A.1 (in Appendix A.1).

**Lemma 4.2.** *For any  $S \in \mathcal{S}$ , we have*

$$|r^*(S) - \bar{r}(S; \mathbf{p}^*)| \leq \epsilon.$$

---

<sup>1</sup>Without loss of generation, we can take  $\epsilon$  such that  $M\epsilon > 1$ .

## 4.2 Converting a Discrete Arm to a Set of Binary Arms

Follow the classical process in [Wang et al. \(2023\)](#), we can convert a discrete arm  $X_i$  to a set of binary arms and estimate the parameters  $\mathbf{q}^* = \{q_{i,j}^* : i \in [N], j \in [M]\}$  instead of  $\mathbf{p}^*$ , where

$$q_{i,j}^* := \frac{p_{i,j}^*}{1 - \sum_{j' > j} p_{i,j'}^*}, \quad p_{i,j}^* = q_{i,j}^* \cdot \prod_{j' > j} (1 - q_{i,j'}^*). \quad (3)$$

Let  $\{\bar{Y}_{i,j}\}_{i \in [N], j \in [M]}$  be independent binary random variables such that  $\bar{Y}_{i,j}$  takes value  $v_j$  with probability  $q_{i,j}^*$ , and value 0 otherwise. Then  $\max_{j \in [M]} \{\bar{Y}_{i,j}\}$  has the same distribution as  $\bar{X}_i$ . For any  $S \in \mathcal{S}$ , define  $\bar{r}_q(S; \mathbf{q})$  as the expected maximum reward of  $\{\bar{Y}_{i,j}\}_{i \in S, j \in [M]}$  with probability set  $\mathbf{q}$ . Then we have

**Lemma 4.3.** *For any  $\mathbf{p}$  and  $\mathbf{q}$  satisfying Eq. (3), we have*

$$\bar{r}_q(S; \mathbf{q}) = \bar{r}(S; \mathbf{p}), \quad \forall S \in \mathcal{S}.$$

The formal version of this lemma is given in Lemma [A.2](#). Moreover, the function  $\bar{r}_q$  is monotone with respect to  $\mathbf{q}$ , i.e.,

**Lemma 4.4** ([Wang et al. \(2023, Lemma 3.1\)](#)). *For two probability set  $\mathbf{q}'$  and  $\mathbf{q}$  such that  $q'_{i,j} \geq q_{i,j}$  holds for any  $i \in [N], j \in [M]$ , we have*

$$\bar{r}_q(S; \mathbf{q}') \geq \bar{r}_q(S; \mathbf{q}), \quad \forall S \in \mathcal{S}.$$

## 4.3 An Efficient Offline Oracle for Discrete $K$ -Max Bandits

For any discrete  $K$ -Max bandits with probability set  $\mathbf{p}$ , we can apply the *PTAS* algorithm ([Chen et al., 2016a](#)) as a polynomial time offline  $\alpha$ -approximation optimization oracle for any given  $\alpha < 1$ . Moreover, for any probability set  $\mathbf{q}$ , we can convert it to  $\mathbf{p}$  by Eq. (3), input this  $\mathbf{p}$  to the PTAS oracle and get the approximation solution  $\text{PTAS}(\mathbf{p})$  satisfying

$$\begin{aligned} \bar{r}_q(\text{PTAS}(\mathbf{p}); \mathbf{q}) &= \bar{r}(\text{PTAS}(\mathbf{p}); \mathbf{p}) \\ &\geq \alpha \cdot \max_{S \in \mathcal{S}} \bar{r}(S; \mathbf{p}) = \alpha \cdot \max_{S \in \mathcal{S}} \bar{r}_q(S; \mathbf{q}). \end{aligned} \quad (4)$$

In the following algorithm, we set  $\alpha = 1 - \epsilon$  and control the relative error to achieve sublinear regret guarantees.

## 4.4 Efficient Algorithm for Continuous $K$ -Max Bandits

Building on the methodology in previous subsections, we adapt the framework in [Wang et al. \(2023\)](#), and present DCK-UCB (Discretized Continuous  $K$ -Max with Upper Confidence Bounds), the first efficient algorithm addressing  $K$ -Max bandits with general continuous outcome distributions. Generally speaking, we first discretize the continuous  $K$ -Max bandits to discrete  $K$ -Max bandits. Then we convert every discrete arm to a set of binary arms, and estimate the corresponding  $\mathbf{q}^*$ . Finally, we convert  $\mathbf{q}^*$  back to  $\mathbf{p}^*$ , input  $\mathbf{p}^*$  to the PTAS oracle, and get the action we want to select.

Algorithm 1 presents the pseudo-code of DCK-UCB. In Line 3, we calculate the optimistic estimator  $\bar{q}_{i,j}^t$  which upper bounds  $q_{i,j}^*$  with high probability. This is done by adding two upper confidence bonus terms  $\beta_{i,j}^t$  and  $(K-1)L^4/j^2$ . Analysis shows that  $\bar{q}_{i,j}^t \geq q_{i,j}^*$  with high probability (Lemma [4.5](#)). The detailed discussion on this estimator will be given in the following

---

**Algorithm 1** DCK-UCB: Discretization Continuous  $K$ -Max Bandits with Upper Confidence Bonus

---

**Input:** Discretization granularity  $\epsilon$ , upper confidence bonuses  $\{\beta_{i,j}^t : i \in [N], j \in [M], t \in [T]\}$ , and the offline  $\alpha$ -approximated optimization oracle PTAS for discrete  $K$ -Max bandits (Chen et al., 2016a).

1: Initialize  $M \leftarrow \lceil 1/\epsilon \rceil$ ,  $\hat{q}_{i,1}^1 \leftarrow 1$  for every  $i \in [N]$ , and  $\hat{q}_{i,j}^1 \leftarrow 0$  for every  $i \in [N], j > 1$ .

2: **for**  $t = 1, 2, \dots, T$  **do**

3:   For every  $i \in [N], j \in [M]$ , set

$$\bar{q}_{i,j}^t \leftarrow \min \left\{ \hat{q}_{i,j}^t + \beta_{i,j}^t + (K-1) \frac{L^4}{j^2}, 1 \right\}. \quad (5)$$

4:   Convert  $\bar{q}^t$  to  $\bar{p}^t$  by Eq. (3).

5:   Choose action  $S_t \leftarrow \text{PTAS}(\bar{p}^t)$ .

6:   Observe  $(r_t, i_t)$  by executing action  $S_t$ . Denote  $j_t$  as the range number of  $r_t$ , i.e.,  $r_t \in M_{j_t}$ .

7:   For any  $i, j \in [N] \times [M]$ ,

$$C_t(i, j) = C_{t-1}(i, j) + \mathbb{1}[i = i_t \ \& \ j = j_t]$$

and

$$SC_t(i, j) = SC_{t-1}(i, j) + \mathbb{1}[i \in S_t \ \& \ j \geq j_t]$$

8:   Calculate estimator  $\hat{q}_{i,j}^{t+1} \leftarrow \frac{C_t(i,j)}{SC_t(i,j)}$ , for every  $i \in [N]$  and  $j \in [M]$ .

9: **end for**

---

paragraphs. In Lines 4-5, the agent converts this  $\bar{q}$  to  $\bar{p}$ , and then runs the offline  $\alpha$ -approximation optimization oracle PTAS with  $\alpha = 1 - \epsilon$  to get action  $S_t$  for execution. In Line 6, the agent gets the value-index return  $(r_t, i_t)$ , and discretizes the value  $r_t$  to the index of bin  $j_t$ , i.e.,  $r_t \in M_{j_t}$ . In Lines 7-8, the agent estimates  $q^*$  by two counters:  $C_t(i, j)$  counts the times when  $(i, j)$  exactly equals the feedback  $(i_t, j_t)$ , and  $SC_t(i, j)$  counts the number of steps  $\tau \leq t$  satisfying  $i \in S_\tau$  and  $j_\tau \leq j$ . As outlined in Algorithm 1, each step of the algorithm has polynomial time and space complexity, which demonstrates the computational tractability of DCK-UCB.

**Biased Estimator.** The key challenge in the algorithm design and theoretical analysis is that  $\hat{q}_{i,j}^t$  is not an unbiased estimator for  $q_{i,j}^*$ . This means that except for the confidence radius due to the randomness of the environment, we still need another bonus term to bound the bias to guarantee that  $\hat{q}_{i,j}^t$  is a UCB for  $q_{i,j}^*$ .

Specifically, note that

$$q_{i,j}^* = \frac{p_{i,j}^*}{1 - \sum_{j' > j} p_{i,j'}^*} = \frac{p_{i,j}^*}{\sum_{j'=1}^j p_{i,j'}^*} = \frac{\mathbb{P}[X_i \in M_j]}{\mathbb{P}[X_i \in M_{\leq j}]}$$

If we have an assumption that when  $i \in S_\tau$  and  $j_\tau = j$ ,  $X_i(\tau) \in M_j$  implies  $i = i_\tau$ , then we can guarantee that  $\hat{q}_{i,j}^t = C_t(i,j)/SC_t(i,j)$  is an unbiased estimator for  $q_{i,j}^*$ . This is because that in this case,  $\frac{C_t(i,j)}{SC_t(i,j)} = \frac{\#\text{ of } i_\tau = i, j_\tau = j}{\#\text{ of } i \in S_\tau, j_\tau \leq j}$  is the fraction of  $X_i(\tau) \in M_j$  condition on  $i \in S_\tau, j_\tau \leq j$ ,

which is an unbiased estimator for

$$\begin{aligned}
& \mathbb{P}[X_i(\tau) \in M_j \mid i \in S_\tau, j_\tau \leq j] \\
&= \frac{\mathbb{P}[X_i(\tau) \in M_j, i \in S_\tau, j_\tau \leq j]}{\mathbb{P}[i \in S_\tau, j_\tau \leq j]} \\
&= \frac{\mathbb{P}[X_i(\tau) \in M_j] \cdot \mathbb{P}[X_k(\tau) \in M_{\leq j}, \forall k \in S_\tau, k \neq i]}{\mathbb{P}[X_i(\tau) \in M_{\leq j}] \cdot \mathbb{P}[X_k(\tau) \in M_{\leq j}, \forall k \in S_\tau, k \neq i]} \\
&= \frac{\mathbb{P}[X_i(\tau) \in M_j]}{\mathbb{P}[X_i(\tau) \in M_{\leq j}]}
\end{aligned}$$

However, we know that in the discrete K-Max bandits converted from the continuous K-Max bandits, there is no such assumption (different from [Wang et al. \(2023\)](#) who requires deterministic tie-breaking rule). When multiple arm has  $X_i(\tau) \in M_j$ , the observed winning arm  $i_t = \arg \max X_i(\tau)$  is not a fixed one, and even we do not know the distribution of the winner. Because of this, we cannot guarantee that condition on  $i \in S_\tau, j_\tau \leq j$ , we increase the counter for every time  $X_i(\tau) \in M_j$ . Some steps that  $X_i(\tau) \in M_j$  but  $X_i(\tau)$  is not the winner are missed. This nondeterministic tie-breaking effect, arising from the continuous-to-discrete transformation, induces systematic negative bias in conventional estimators  $\{\hat{q}_{i,j}^t\}$ . Therefore, to guarantee that our used  $\{\hat{q}_{i,j}^t\}$  is an upper confidence bound of  $\{q_{i,j}^*\}$ , we need another bonus term (i.e., the term  $(K-1)\frac{L^4}{j^2}$ ), given by a novel concentration analysis with bias-aware error control. This is shown in the following key lemma, where the formal version is in Lemma [A.3](#).

**Lemma 4.5.** *Under Assumption 4.1, let the confidence radius be defined as*

$$\beta_{i,j}^t := \sqrt{8 \frac{\log(NMt)}{SC_{t-1}(i,j)}}. \quad (6)$$

Then with probability at least  $1 - t^{-2}$ ,

$$|\hat{q}_{i,j}^t - q_{i,j}^*| \leq \beta_{i,j}^t + (K-1) \cdot (L^4/j^2), \quad (7)$$

holds for every  $t \in [T]$ ,  $i \in [N]$  and  $j \in [M]$ .

The bound in Lemma 4.5 decomposes into an exploration bonus term  $\beta_{i,j}^t$  and a bias compensation term  $(K-1)\frac{L^4}{j^2}$ . The exploration bonus term arises from the randomness of the environment, which is almost the same with existing researches ([Wang and Chen, 2017](#); [Liu et al., 2023b](#); [Wang et al., 2023](#)). The bias compensation term, on the other hand, comes from the nondeterministic tie-breaking effect in the continuous-to-discrete transformation. As we have explained, this term is because that condition on  $i \in S_\tau, j_\tau \leq j$ , there are some time steps that  $X_i(\tau) \in M_j$  but arm  $i$  is not the winner and thus we miss these steps in counter  $C_{i,j}^t$ . When this happens, we know that there must be at least one other arm  $i' \neq i$ ,  $i' \in S_\tau$  such that  $X_{i'}(\tau) \in M_j$ . This probability can be upper bounded by

$$\begin{aligned}
& \sum_{i' \neq i, i' \in S_\tau} \mathbb{P}[X_i(\tau) \in M_j, X_{i'}(\tau) \in M_j \mid i \in S_\tau, j_\tau \leq j] \\
&= \sum_{i' \neq i, i' \in S_\tau} \frac{p_{i,j}^* p_{i',j}^*}{\sum_{j' \leq j} p_{i,j'}^* \sum_{j' \leq j} p_{i',j'}^*} \leq (K-1) \frac{(L\epsilon)^2}{(j\epsilon/L)^2},
\end{aligned}$$

where the last inequality is because of bi-Lipschitz assumption Assumption 4.1.

Notably, the bias term dominates for small  $j$  values due to the influence of other arms becomes higher when condition on  $j_\tau \leq j$  with smaller  $j$ . However, our regret analysis in Section 4.5 suggests that the amplified bias for small  $j$  has diminishing impact on cumulative regret – a crucial property enabling our sublinear regret guarantee.

## 4.5 Theoretical Results

We establish the first efficient algorithm DCK-UCB (Algorithm 1) which enjoys the sublinear regret guarantees in continuous  $K$ -Max bandits problem with value-index feedback.

**Theorem 4.6.** *Under Assumption 4.1, let the offline optimization oracle be a PTAS implementation (Chen et al., 2016a). Given the exploration bonus term  $\beta_{i,j}^t$  in Eq. (6), discretization granularity  $\epsilon = \mathcal{O}(L^{-2}K^{-3/4}N^{1/4}T^{-1/4})$  and PTAS approximation factor  $\alpha = 1 - \epsilon$ , Algorithm 1 enjoys the regret guarantee*

$$\mathcal{R}(T) \leq \tilde{\mathcal{O}}(L^2 N^{\frac{1}{4}} K^{\frac{5}{4}} T^{\frac{3}{4}}).$$

The formal statement with precise constants appears in Theorem A.13. Our analysis reveals that careful calibration of the discretization-error versus statistical-estimation trade-off enables the first sublinear regret guarantee  $\mathcal{O}(T^{3/4})$  for continuous  $K$ -Max bandits.

**Comparison to Prior Works.** The  $\mathcal{O}(T^{3/4})$  regret upper bound of DCK-UCB (Algorithm 1) shown in Theorem 4.6 advances the state-of-the-art in several directions. Wang et al. (2023) can achieve an  $\mathcal{O}(\sqrt{T})$  regret upper bound in the discrete  $K$ -Max bandits, but their algorithm cannot work for the continuous case due to non-zero discretization error and nondeterministic tie-breaking. Recent work on submodular bandits (Pasteris et al., 2023; Fourati et al., 2024) attains  $O(T^{2/3})$  regret via greedy oracles, but this approach suffers dual limitations: (1) The baseline of their regret is  $\sum_{t=1}^T (1 - 1/e)r^*(S^*)$ , but not  $\sum_{t=1}^T r^*(S^*)$ . In our definition, their regret becomes linear. (2) Their algorithm requires the availability of submitting any subset of  $\mathcal{A}$  with size less than or equal to  $K$ , which may not be practical in some applications, such as recommendation systems or portfolio selection that need to always submit size  $K$  subsets. Our framework resolves both issues through our novel bias-corrected estimators with PTAS integration, which is both efficient and effective in dealing with continuous  $K$ -Max bandits.

## 4.6 Proof Sketch of Theorem 4.6

In this section we outline the proof of Theorem 4.6, which consists of four main steps.

**Step 1: From continuous regret to discretized regret.** Let  $\Delta_t := r^*(S^*) - r^*(S_t)$  be the regret for each round  $t$ . To control the regret, we aim to bound the summation of  $\Delta_t$ .

$$\mathcal{R}(T) = \mathbb{E} \left[ \sum_{t=1}^T \Delta_t \right].$$

We first transfer the regret from continuous  $K$ -Max bandits to the discrete case. With Lemma 4.2, we have

$$\Delta_t \leq \bar{r}(S^*; \mathbf{p}^*) - \bar{r}(S_t; \mathbf{p}^*) + 2\epsilon.$$

**Step 2: From discretized regret to estimation error.** Recall the definition of  $\bar{r}_q$  in Section 4.2, we have  $\bar{r}_q(S; \mathbf{q}) = \bar{r}(S; \mathbf{p})$  for any  $S \in \mathcal{S}$ , and probability set  $\mathbf{p}, \mathbf{q}$  satisfying Eq. (3). By the monotonicity of  $\bar{r}_q$  (in Lemma 4.4) and the concentration analysis (in Lemma 4.5), we have with high probability,  $\forall (i, j) \in [N] \times [M]$ ,  $\bar{q}_{i,j}^t \geq q_{i,j}^*$  holds for any  $t \in [T]$ , which implies

$$\bar{r}_q(S^*; \bar{\mathbf{q}}^t) \geq \bar{r}_q(S^*; \mathbf{q}^*).$$

Moreover, by the property of  $\alpha$ -approximated offline optimization oracle PTAS (Chen et al., 2016a) (in Eq. (4)) with  $\alpha = 1 - \epsilon$ ,

$$(1 - \epsilon)\bar{r}_q(S^*; \bar{\mathbf{q}}^t) \leq (1 - \epsilon) \max_{S \in \mathcal{S}} \bar{r}_q(S; \bar{\mathbf{q}}^t) \leq \bar{r}_q(S_t; \bar{\mathbf{q}}^t),$$

which implies the conversion from  $\Delta_t$  to the estimation error term

$$\begin{aligned} \Delta_t &\leq \bar{r}_q(S^*; \bar{\mathbf{q}}^t) - \bar{r}_q(S_t; \mathbf{q}^*) + 2\epsilon \\ &\leq (1 - \epsilon)\bar{r}_q(S^*; \bar{\mathbf{q}}^t) - \bar{r}_q(S_t; \mathbf{a}^*) + 3\epsilon \\ &\leq \bar{r}_q(S_t; \bar{\mathbf{q}}^t) - \bar{r}_q(S_t; \mathbf{q}^*) + 3\epsilon. \end{aligned}$$

Therefore, we then focus on bounding the estimation error  $\bar{\Delta}_t := \bar{r}_q(S_t; \bar{\mathbf{q}}^t) - \bar{r}_q(S_t; \mathbf{q}^*)$  to guarantee the sublinear regret upper bound:

$$\mathcal{R}(T) = \mathbb{E} \left[ \sum_{t=1}^T \Delta_t \right] \leq \mathbb{E} \left[ \sum_{t=1}^T \bar{\Delta}_t \right] + 3T\epsilon. \quad (8)$$

**Step 3: Decompose the estimation error.** By similar methods as achieving the Triggering Probability Modulated (TPM) smoothness condition in cascading bandits (Wang and Chen, 2017) and  $K$ -Max bandits for binary distributions (Wang et al., 2023), we propose the following lemma.

**Lemma 4.7.** Denote the probability of event  $\{j_t \leq j\}$  as

$$Q_j^*(S_t) := \prod_{k \in S_t, j' > j} (1 - q_{k,j'}^*).$$

Then we have

$$\bar{\Delta}_t \leq 2 \sum_{i \in S_t, j \in [M]} Q_j^*(S_t) \cdot v_j \cdot |\bar{q}_{i,j}^t - q_{i,j}^*|. \quad (9)$$

Equipped with Lemma 4.7, we decompose  $\bar{\Delta}_t$  into two parts through our novel concentration analysis in Lemma 4.5 and the definition of optimistic estimator  $\bar{q}_{i,j}^t$  in Eq. (5)

$$\begin{aligned} \bar{\Delta}_t &\leq \underbrace{4 \sum_{i \in S_t, j \in [M]} Q_j^*(S_t) \cdot v_j \cdot \beta_{i,j}^t}_{\text{Bonus}_t} \\ &\quad + \underbrace{4 \sum_{i \in S_t, j \in [M]} Q_j^*(S_t) \cdot v_j \cdot (K-1) \frac{L^4}{j^2}}_{\text{Bias}_t}. \end{aligned} \quad (10)$$

**Step 4: Bound the Bonus and Bias terms.** For the Bonus term, we apply standard analysis for combinatorial bandits with triggering arms in (Wang and Chen, 2017; Liu et al., 2023b) where we encounter  $NM$  binary arms in total and select  $KM$  binary arms in every action and get

$$\mathbb{E} \left[ \sum_{t=1}^T \text{Bonus}_t \right] \leq \tilde{\mathcal{O}} \left( \sqrt{(NM) \cdot (KM) \cdot T} \right). \quad (11)$$

To control the bias terms, we recall that  $v_j = (j - 1)\epsilon$ . Therefore, we can write

$$\begin{aligned} \text{Bias}_t &\leq 4K^2L^4 \sum_{j \in [M]} (j - 1)\epsilon / j^2 \\ &\leq \mathcal{O}(K^2L^4\epsilon \log(M)), \end{aligned} \quad (12)$$

Therefore, combining Eqs. (8) and (10) to (12), the regret can be bounded by

$$\begin{aligned} \mathcal{R}(T) &\leq \mathbb{E} \left[ \sum_{t=1}^T \text{Bonus}_t + \text{Bias}_t \right] + \mathcal{O}(T\epsilon) \\ &\leq \tilde{\mathcal{O}} \left( \sqrt{NKM^2T} + K^2L^4T\epsilon \right), \end{aligned}$$

where  $M = \lceil 1/\epsilon \rceil$ . By taking  $\epsilon = \mathcal{O}(T^{-\frac{1}{4}}K^{-\frac{3}{4}}N^{\frac{1}{4}}L^{-2})$ , we have

$$\mathcal{R}(T) \leq \tilde{\mathcal{O}} \left( L^2N^{\frac{1}{4}}K^{\frac{5}{4}}T^{\frac{3}{4}} \right).$$

## 5 Better Performance in a Special Case: Exponential Distributions

In this section, we demonstrate how specific distributional structure enables the improvement of the regret guarantee from  $\tilde{\mathcal{O}}(T^{\frac{3}{4}})$  to  $\tilde{\mathcal{O}}(\sqrt{T})$ . Specifically, we investigate the special case where each distribution  $D_i$  for  $i \in [N]$  follows the exponential distribution with linear parameterization.

Exponential distributions naturally model arrival or failure times in networked systems, job completion times in distributed computing, and service durations in queuing systems. A canonical application arises in server scheduling, where the goal is to select  $K$  servers to minimize the service latency. Here, each server's latency can be modeled as an exponential random variable with a rate parameter  $\mu_i$ , and the overall performance of the  $K$  selected servers is the lowest latency achieved among them. Here, the random outcome  $X_i$  can be viewed as a random loss, and the winning loss is the minimum one. Moreover, we consider a linear parameterization to parameter  $\mu_i$ , which allows incorporating features like distance, traffic, or weather conditions into the model.

### 5.1 The $K$ -Min Exponential Bandits

Based on the intuition, in this section we consider a special case of  $K$ -Max bandits: the  $K$ -Min exponential bandits. Here each arm  $i$  generates loss  $X_i$  from an exponential distribution with linear parameterization. Specifically, each outcome distribution is an exponential distribution, i.e.,  $X_i \sim D_i = \text{Exp}(\mu_i)$  where  $\mu_i > 0$  is the parameter of arm  $i$ . Moreover, we assume that there exists a  $d$ -dimension unknown parameter  $\theta^* \in \mathbb{R}^d$  and a known feature mapping  $\phi : [N] \rightarrow \mathbb{R}^d$  such that  $\mu_i = \langle \phi(i), \theta^* \rangle$  holds for any  $i \in [N]$ . The feature mapping  $\phi$  satisfies that  $\|\phi(i)\|_2 \leq 1$

and the unknown parameter  $\theta^*$  satisfies  $\theta^* \in \Theta \subset \mathbb{R}^d$ , where  $\sup_{\theta \in \Theta} \|\theta\|_2 \leq V$ . The agent observes *only* the minimum loss  $\ell_t = \min_{i \in S_t} X_i(t)$  after playing subset  $S_t \in \mathcal{S} = \{S \subseteq [N] : |S| = K\}$ . That is, we consider the weaker full bandit feedback case.

Let  $\ell^*(S) := \mathbb{E}[\ell_t | S]$  be the expected loss for action  $S \in \mathcal{S}$ , we further denote the best action  $S^* = \operatorname{argmin}_{S \in \mathcal{S}} \ell^*(S)$  and similarly introduce the regret metric to evaluate the performance of this agent:

$$\mathcal{R}(T) = \mathbb{E} \left[ \sum_{t=1}^T \ell^*(S_t) - \ell^*(S^*) \right].$$

Note that we can let  $Z_i(t) = -X_i(t)$  and view  $Z_i(t)$  as a kind of reward, and let  $r_t = \max_{i \in S_t} Z_i(t)$ . Then we can see that  $\ell_t = \min_{i \in S_t} X_i(t) = \min_{i \in S_t} -Z_i(t) = -\max_{i \in S_t} Z_i(t) = -r_t$ . By this way, we can view  $K$ -Min exponential bandits as a special case of  $K$ -Max bandits. However, one important difference is that in  $K$ -Min exponential bandits, we do not have value-index feedback, i.e., we do not know the winner's index. This is a full *bandit feedback* setting, and making  $K$ -Min exponential bandits even more challenging.

## 5.2 Algorithm and Results

The key observation in  $K$ -Min exponential bandits is that the minimum of several exponential distributions still follows an exponential distribution. That is, we have

$$\min_{i \in S} X_i \sim \text{Exp} \left( \sum_{i \in S} \mu_i \right) = \text{Exp} \left( \sum_{i \in S} \langle \phi(i), \theta^* \rangle \right).$$

Therefore, it becomes much easier to estimate the true parameter  $\theta^*$  by MLE. Specifically, let  $\psi(S) := \sum_{i \in S} \phi(i)$ ,  $\forall S \in \mathcal{S}$ . Then with chosen action  $S_t$  and parameter  $\theta$ , the observed loss should follow the exponential distribution  $\text{Exp}(\sum_{i \in S} \phi(i)^T \theta) = \text{Exp}(\psi(S)^T \theta)$ , whose probability density function is  $f(x) = \psi(S)^T \theta e^{(-\psi(S)^T \theta x)}$ . Because of this, the log-likelihood function is

$$L_t(\ell_t; S_t, \theta) := -\log \left( \psi(S_t)^T \theta e^{(-\psi(S_t)^T \theta \ell_t)} \right). \quad (13)$$

Denote  $\mathcal{L}_t(\theta)$  as the summation of  $L_t$  and a regularization term

$$\mathcal{L}_t(\theta; \lambda) := \sum_{i < t} L_i(\ell_i; S_i, \theta) + \frac{\lambda}{2} \|\theta\|^2, \quad (14)$$

where  $\lambda$  is the regularization factor. Then we present the algorithm **MLE-Exp** for  $K$ -Min exponential bandits in Algorithm 2.

In Line 2 of Algorithm 2, we estimate the MLE  $\hat{\theta}_t$  by minimizing the summation of the log-likelihood function and the regularization term  $\mathcal{L}_t(\theta, \lambda_t)$ . Given  $\lambda_t$  a priori, we will write  $\mathcal{L}_t(\theta)$  instead of  $\mathcal{L}_t(\theta, \lambda_t)$  for simplicity. Inspired by Liu et al. (2024a); Lee et al. (2024); Liu et al. (2024b), in Line 3, we construct a confidence set  $C_t(\hat{\theta}_t; \delta)$ , centered at the MLE  $\hat{\theta}_t$  with confidence radius  $\gamma_t(\delta)$ , based on the gradient term  $g_t(\theta) := -\nabla_\theta \mathcal{L}_t(\theta) + \sum_{i < t} \ell_i \psi(S_i)$  and Hessian matrix  $H_t(\theta) := \nabla_\theta^2 \mathcal{L}_t(\theta)$ :

$$C_t(\hat{\theta}_t; \delta) := \left\{ \theta \in \Theta : \|g_t(\theta) - g_t(\hat{\theta}_t)\|_{H_t^{-1}(\theta)} \leq \gamma_t(\delta) \right\}, \quad (15)$$

---

**Algorithm 2** MLE-Exp: MLE for  $K$ -Min Exponential Bandits

---

**Input:** Regularization factors  $\{\lambda_t\}_{t \in [T]}$ , confidence radius  $\{\gamma_t\}_{t \in [T]}$ , and probability constant  $\delta$ .

- 1: **for**  $t = 1, \dots, T$  **do**
- 2:   Compute MLE  $\hat{\theta}_t$  by

$$\hat{\theta}_t \leftarrow \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \mathcal{L}_t(\theta; \lambda_t),$$

where  $\mathcal{L}_t(\theta; \lambda)$  is given in Eq. (14).

- 3:   Construct the confidence set  $C_t(\hat{\theta}_t; \delta, \lambda_t)$  according to Eq. (29)
- 4:    $(S_t, \tilde{\theta}_t) \leftarrow \underset{S \in \mathcal{S}, \theta \in C_t(\hat{\theta}_t; \delta, \lambda_t)}{\operatorname{argmax}} \langle \psi(S), \theta \rangle$
- 5:   Play action  $S_t$  and observe the loss  $\ell_t$ .

- 6: **end for**
- 

where  $\gamma_t$  is the confidence radius. Then in Line 4, we apply a double oracle to look for the action  $S$  whose expected loss under a parameter  $\theta$  in the confidence set  $(1/\langle \psi(S), \theta \rangle)$  is minimized. Finally, we select this greedy action in Line 5 and use the observation to update the next time step's MLE and confidence set.

The regret guarantees of Algorithm 2 is given below.

**Theorem 5.1.** *With  $\delta = 1/T$ ,  $\lambda_t = \Theta(d \log T)$ , and  $\gamma_t = \Theta(\sqrt{d \log T})$ , Algorithm 2 satisfies:*

$$\mathcal{R}(T) \leq \tilde{\mathcal{O}}\left(\sqrt{d^3 T}\right).$$

Compared with the  $O(T^{3/4})$  regret upper bound for general continuous K-Max bandits, here the regret upper bound is reduced to  $O(T^{1/2})$  (which is nearly minimax optimal) even without the feedback of winner's index, due to the utilization of the exponential distribution's property. In short, we do not need to use a discretization method and can directly construct an unbiased estimator for the known parameter  $\theta^*$ . The proof is inspired by previous analysis of general linear bandits (Lee et al., 2024; Liu et al., 2024a) and logistics bandits (Liu et al., 2024b), and we defer the detailed proof to Appendix B.

## 6 Conclusion and Future Work

We presented the first computationally efficient algorithm DCK-UCB (Algorithm 1) for Continuous  $K$ -Max Bandits with value-index feedback, resolving fundamental challenges in handling general continuous outcome distributions in  $K$ -Max Bandits and achieving the first sublinear regret guarantees  $\tilde{\mathcal{O}}(T^{3/4})$ . When considering exponential distributions as a special case of continuous  $K$ -Max bandits, we demonstrated that an MLE-based algorithm MLE-Exp (Algorithm 2) can achieve the  $\tilde{\mathcal{O}}(\sqrt{T})$  regret upper bound (Theorem 5.1) even under full-bandit feedback, which further advances the general result.

Further enhancing the  $\tilde{\mathcal{O}}(T^{3/4})$  regret for the general continuous distribution case is an interesting future direction. One potential avenue involves developing variance-aware algorithms that adapt to second-order statistics of the outcomes. Such methods might theoretically reduce the regret to  $\tilde{\mathcal{O}}(T^{2/3})$  through refined analysis for the variance-adaptive exploration bonus terms, inspired by its successful applications in CMABs (Liu et al., 2023b, 2024b). However, such approaches face inherent challenges due to the biased estimations induced by nondeterministic tie-breaking, which create new concentration challenges for variance terms of biased observations. Overcoming these limitations may require developing new bias-corrected concentrations

for variance estimators or alternative feedback models tailored to continuous outcomes. Other directions include developing lower bounds and relaxing the bi-Lipschitz assumption.

## References

- Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24.
- Agarwal, A., Johnson, N., and Agarwal, S. (2020). Choice bandits. *Advances in neural information processing systems*, 33:18399–18410.
- Cesa-Bianchi, N. and Lugosi, G. (2012). Combinatorial bandits. *Journal of Computer and System Sciences*, 78(5):1404–1422.
- Chen, S., Lin, T., King, I., Lyu, M. R., and Chen, W. (2014). Combinatorial pure exploration of multi-armed bandits. *Advances in neural information processing systems*, 27.
- Chen, W., Hu, W., Li, F., Li, J., Liu, Y., and Lu, P. (2016a). Combinatorial multi-armed bandit with general reward functions. *Advances in Neural Information Processing Systems*, 29.
- Chen, W., Wang, Y., and Yang, S. (2009). Efficient influence maximization in social networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 199–208.
- Chen, W., Wang, Y., and Yuan, Y. (2013). Combinatorial multi-armed bandit: General framework and applications. In *International conference on machine learning*, pages 151–159. PMLR.
- Chen, W., Wang, Y., Yuan, Y., and Wang, Q. (2016b). Combinatorial multi-armed bandit and its extension to probabilistically triggered arms. *Journal of Machine Learning Research*, 17(50):1–33.
- Combes, R., Talebi Mazraeh Shahi, M. S., Proutiere, A., et al. (2015). Combinatorial bandits revisited. *Advances in neural information processing systems*, 28.
- Fourati, F., Quinn, C. J., Alouini, M.-S., and Aggarwal, V. (2024). Combinatorial stochastic-greedy bandit. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 12052–12060.
- Gai, Y., Krishnamachari, B., and Jain, R. (2012). Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations. *IEEE/ACM Transactions on Networking*, 20(5):1466–1478.
- Goel, A., Guha, S., and Munagala, K. (2006). Asking the right questions: Model-driven optimization using probes. In *Proceedings of the twenty-fifth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 203–212.
- Gopalan, A., Mannor, S., and Mansour, Y. (2014). Thompson sampling for complex online problems. In *International conference on machine learning*, pages 100–108. PMLR.
- Janz, D., Liu, S., Ayoub, A., and Szepesvári, C. (2024). Exploration via linearly perturbed loss minimisation. In *International Conference on Artificial Intelligence and Statistics*, pages 721–729. PMLR.

- Jun, K.-S., Jamieson, K., Nowak, R., and Zhu, X. (2016). Top arm identification in multi-armed bandits with batch arm pulls. In *Artificial Intelligence and Statistics*, pages 139–148. PMLR.
- Kveton, B., Wen, Z., Ashkan, A., and Szepesvari, C. (2015a). Combinatorial cascading bandits. *Advances in Neural Information Processing Systems*, 28.
- Kveton, B., Wen, Z., Ashkan, A., and Szepesvari, C. (2015b). Tight regret bounds for stochastic combinatorial semi-bandits. In *Artificial Intelligence and Statistics*, pages 535–543. PMLR.
- Lattimore, T. and Szepesvári, C. (2020). *Bandit algorithms*. Cambridge University Press.
- Lee, J., Yun, S.-Y., and Jun, K.-S. (2024). A unified confidence sequence for generalized linear models, with applications to bandits. *arXiv preprint arXiv:2407.13977*.
- Li, L., Lu, Y., and Zhou, D. (2017). Provably optimal algorithms for generalized linear contextual bandits. In *International Conference on Machine Learning*, pages 2071–2080. PMLR.
- Liu, Q., Netrapalli, P., Szepesvari, C., and Jin, C. (2023a). Optimistic mle: A generic model-based algorithm for partially observable sequential decision making. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 363–376.
- Liu, S., Ayoub, A., Sentenac, F., Tan, X., and Szepesvári, C. (2024a). Almost free: Self-concordance in natural exponential families and an application to bandits. *arXiv preprint arXiv:2410.01112*.
- Liu, X., Dai, X., Wang, X., Hajiesmaili, M., and Lui, J. (2024b). Combinatorial logistic bandits. *arXiv preprint arXiv:2410.17075*.
- Liu, X., Zuo, J., Wang, S., Lui, J. C., Hajiesmaili, M., Wierman, A., and Chen, W. (2023b). Contextual combinatorial bandits with probabilistically triggered arms. In *International Conference on Machine Learning*, pages 22559–22593. PMLR.
- Nemhauser, G. L., Wolsey, L. A., and Fisher, M. L. (1978). An analysis of approximations for maximizing submodular set functions—i. *Mathematical programming*, 14:265–294.
- Nie, G., Agarwal, M., Umrawal, A. K., Aggarwal, V., and Quinn, C. J. (2022). An explore-then-commit algorithm for submodular maximization under full-bandit feedback. In *Uncertainty in Artificial Intelligence*, pages 1541–1551. PMLR.
- Pasteris, S., Rumi, A., Vitale, F., and Cesa-Bianchi, N. (2023). Sum-max submodular bandits. *arXiv preprint arXiv:2311.05975*.
- Simchowitz, M., Jamieson, K., and Recht, B. (2016). Best-of-k-bandits. In *Conference on Learning Theory*, pages 1440–1489. PMLR.
- Slivkins, A. et al. (2019). Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286.
- Streeter, M. and Golovin, D. (2008). An online algorithm for maximizing submodular functions. *Advances in Neural Information Processing Systems*, 21.
- Upfal, E. and Mitzenmacher, M. (2005). Probability and computing.
- Wang, Q. and Chen, W. (2017). Improving regret bounds for combinatorial semi-bandits with probabilistically triggered arms and its applications. *Advances in Neural Information Processing Systems*, 30.

- Wang, Y., Chen, W., and Vojnović, M. (2023). Combinatorial bandits for maximum value reward function under max value-index feedback. *arXiv preprint arXiv:2305.16074*.
- Wang, Y., Wang, R., Du, S. S., and Krishnamurthy, A. (2019). Optimism in reinforcement learning with generalized linear function approximation. *arXiv preprint arXiv:1912.04136*.
- Yue, Y. and Guestrin, C. (2011). Linear submodular bandits and their application to diversified retrieval. *Advances in Neural Information Processing Systems*, 24.

# Appendices

<b>A Omitted Proofs in Section 4</b>	<b>18</b>
A.1 Discretization Error . . . . .	18
A.2 Converting to Binary Arms . . . . .	19
A.3 Biased Concentration . . . . .	20
A.4 Optimistic Estimation . . . . .	22
A.5 Regret Decomposition . . . . .	22
A.6 Bounding the Bonus Terms . . . . .	24
A.7 Bounding the Bias Terms . . . . .	27
A.8 Proof of Theorem 4.6 . . . . .	28
<b>B Omitted proofs in Section 5</b>	<b>29</b>
B.1 Concentration Argument for MLE . . . . .	29
B.2 Proof of Theorem 5.1 . . . . .	31

## A Omitted Proofs in Section 4

In this section, we present the omitted proofs in Section 4, which include the full proof of Theorem 4.6.

### A.1 Discretization Error

First we show that the discretization from original continuous problem  $\mathcal{B}^*$  to  $\bar{\mathcal{B}}$  with discretization width  $\epsilon$  will involve controllable error in expected loss, which is shown in Lemma 4.2 and formalized by the following lemma.

**Lemma A.1.** *For any  $S \in \mathcal{S}$ , we have*

$$\bar{r}(S; \mathbf{p}^*) \leq r^*(S) \leq \bar{r}(S; \mathbf{p}^*) + \epsilon. \quad (16)$$

*Proof.* Notice that we have

$$\begin{aligned} \mathbb{P}\left[\max_{i \in S}(\bar{X}_i) = v_j\right] &= \sum_{I \subset S} \prod_{i \in I} \mathbb{P}[\bar{X}_i = v_j] \cdot \prod_{k \in S, k \notin I} \mathbb{P}[\bar{X}_k < v_j] \\ &= \sum_{I \subset S} \prod_{i \in I} \mathbb{P}[X_i \in M_j] \cdot \prod_{k \in S, k \notin I} \mathbb{P}[X_k \in M_{\leq j-1}] \\ &= \mathbb{P}\left[\max_{i \in S}(X_i) \in M_j\right]. \end{aligned}$$

Therefore, by definition of  $r^*(S)$ , we have

$$\begin{aligned} r^*(S) &= \sum_{j \in [M]} \int_{r \in M_j} r \cdot d\mathbb{P}_{\max_{i \in S}(X_i)}(r) \\ &\geq \sum_{j \in [M]} (j-1)\epsilon \int_{r \in M_j} d\mathbb{P}_{\max_{i \in S}(X_i)}(r) \\ &= \sum_{j \in [M]} (j-1)\epsilon \cdot \mathbb{P}\left[\max_{i \in S}(X_i) \in M_j\right] \\ &= \sum_{j \in [M]} (j-1)\epsilon \cdot \mathbb{P}\left[\max_{i \in S}(\bar{X}_i) = (j-1)\epsilon\right] \\ &= \bar{r}(S; \mathbf{p}^*), \end{aligned}$$

where the inequality is given by the definition of  $M_j$ . Then we achieve the left-hand side of Eq. (16). For the other side, we can similarly establish

$$\begin{aligned} r^*(S) &= \sum_{j \in [M]} \int_{r \in M_j} r \cdot d\mathbb{P}_{\max_{i \in S}(X_i)}(r) \\ &\leq \sum_{j \in [M]} j\epsilon \int_{r \in M_j} d\mathbb{P}_{\max_{i \in S}(X_i)}(r) \\ &= \sum_{j \in [M]} (j-1)\epsilon \cdot \mathbb{P}\left[\max_{i \in S}(X_i) \in M_j\right] + \epsilon \cdot \sum_{j \in [M]} \mathbb{P}\left[\max_{i \in S}(X_i) \in M_j\right] \\ &= \bar{r}(S; \mathbf{p}^*) + \epsilon. \end{aligned}$$

□

## A.2 Converting to Binary Arms

As detailed in Section 4.2, we set

$$q_{i,j}^* := \frac{p_{i,j}^*}{1 - \sum_{j' > j} p_{i,j'}^*}, \quad p_{i,j}^* = q_{i,j}^* \cdot \prod_{j' > j} (1 - q_{i,j'}^*),$$

which implies

$$q_{i,j}^* = \frac{p_{i,j}^*}{1 - \sum_{j' > j} p_{i,j'}^*} = \frac{p_{i,j}^*}{\sum_{j'=1}^j p_{i,j'}^*} = \frac{\mathbb{P}[X_i \in M_j]}{\mathbb{P}[X_i \in M_{\leq j}]}.$$

For any given probability set  $\mathbf{q} = \{q_{i,j} : i \in [N], j \in [M]\}$ , we can apply Eq. (3) to get the corresponding  $\mathbf{p}$  defined as

$$p_{i,j} = q_{i,j} \cdot \prod_{j' > j} (1 - q_{i,j'}).$$

Assume  $\{Y_{i,j}^{\mathbf{q}}\}_{i \in [N], j \in [M]}$  is the set of independent binary random variables that  $Y_{i,j}^{\mathbf{q}}$  takes value  $v_j = (j-1)\epsilon$  with probability  $q_{i,j}$  and takes value 0 otherwise. And  $\{X_i^{\mathbf{p}}\}_{i \in [N]}$  is the set of independent discrete random variables that  $X_i^{\mathbf{p}}$  takes value  $v_j$  with probability  $p_{i,j}$  for every  $j \in [M]$ . Therefore, by simple calculation, we have  $\max_{j \in [M]} \{Y_{i,j}^{\mathbf{q}}\}$  has the same distribution of  $X_i^{\mathbf{p}}$ .

$\bar{r}_q(S; \mathbf{q})$  is defined as the expected maximum reward of  $\{Y_{i,j}^{\mathbf{q}}\}_{i \in S, j \in [M]}$ . Then we can write

$$\bar{r}_q(S; \mathbf{q}) = \sum_{j \in [M]} v_j \cdot (Q_j(S; \mathbf{q}) - Q_{j-1}(S; \mathbf{q})), \quad (17)$$

where we denote for simplicity

$$Q_j(S; \mathbf{q}) := \prod_{k \in S, j' > j} (1 - q_{k,j'}). \quad (18)$$

$Q_j(S; \mathbf{q})$  is actually the probability of the event that every arm in  $\{\bar{Y}_{k,j'}\}_{k \in S, j' > j}$  does not sample a non-zero value.

Equipped with the above statement, we can establish the following lemma:

**Lemma A.2.** *For any  $\mathbf{p}$  and  $\mathbf{q}$  satisfying Eq. (3), we have for any  $S \in \mathcal{S}$ ,*

$$\bar{r}_q(S; \mathbf{q}) = \bar{r}(S; \mathbf{p}).$$

*Proof.* Notice that by definition, we have

$$\bar{r}(S; \mathbf{p}) = \mathbb{E} \left[ \max_{i \in S} X_i^{\mathbf{p}} \right],$$

and

$$\bar{r}_q(S; \mathbf{q}) = \mathbb{E} \left[ \max_{i \in S} \max_{j \in [M]} Y_{i,j}^{\mathbf{q}} \right].$$

Notice that  $\max_{j \in [M]} \{Y_{i,j}^{\mathbf{q}}\}$  has the same distribution of  $X_i^{\mathbf{p}}$ , we have

$$\begin{aligned} \bar{r}_q(S; \mathbf{q}) &= \mathbb{E} \left[ \max_{i \in S} \max_{j \in [M]} Y_{i,j}^{\mathbf{q}} \right] \\ &= \mathbb{E} \left[ \max_{i \in S} X_i^{\mathbf{p}} \right] = \bar{r}(S; \mathbf{p}). \end{aligned}$$

□

### A.3 Biased Concentration

We aim to use  $\hat{q}_{i,j}^t$  to estimate  $q_{i,j}^*$ . However, this is a biased estimation. In this section, we carefully control the gap between the biased estimator  $\hat{q}_{i,j}^t$  and the true probability  $q_{i,j}^*$ .

We set  $c_t(i,j) := \mathbb{1}[(i_t, j_t) = (i, j)]$  which is  $\mathcal{F}_t$ -measurable. Then Algorithm 1 counts the summation of  $c_t(i,j)$  as  $C_t(i,j)$ :

$$C_t(i,j) = \sum_{\tau=1}^t c_\tau(i,j),$$

which is  $\mathcal{F}_{t-1}$ -measurable.

For given action  $S_t$  in round  $t$ , the environment will sample a set of outcomes  $\{X_i(t) \sim D_i : i \in S_t\}$ . The value-index feedback is  $r_t = \max_{i \in S_t} X_i(t)$ ,  $i_t = \operatorname{argmax}_{i \in S_t} X_i(t)$ . Algorithm 1 consider  $j_t$  such that  $r_t \in M_{j_t}$ . We denote  $I_t = \operatorname{argmax}_{i \in S_t} \bar{X}_i(t)$ , where  $\bar{X}_i(t)$  is the discretized of  $X_i(t)$  induced by Eq. (2). Notice that under event  $\mathcal{E}_0$ ,  $\operatorname{argmax}_{i \in S} X_i(t)$  is unique. But  $I_t$  might be a set with multiple indices. We emphasize that  $S_t$  is  $\mathcal{F}_{t-1}$  measurable and  $(i_t, r_t, j_t, I_t)$  are  $\mathcal{F}_t$  measurable.

Then we can provide the following lemma.

**Lemma A.3.** *Under event  $\mathcal{E}_0$ , we have for every  $t \in [T]$  and  $(i, j) \in [N] \times [M]$ ,*

$$|\hat{q}_{i,j}^t - q_{i,j}^*| \leq \sqrt{8 \frac{\log(NMt)}{SC_t(i,j)}} + (K-1) \cdot (L^4/j^2),$$

with probability at least  $1 - T^{-2}$ , where we denote this good event as  $\mathcal{E}_1$ .

*Proof.* Denote  $q_{i,j}(S_t) := \mathbb{1}[i \in S_t] \cdot \mathbb{P}[(i_t, j_t) = (i, j) | j_t \leq j, S_t]$ , and  $q_{i,j}^*(S_t) := \mathbb{1}[i \in S_t] \cdot \mathbb{P}[I_t \ni i, j_t = j | j_t \leq j, S_t]$ . Therefore, for given  $i \in [N], j \in [M]$ , we have

$$\mathbb{E}[\mathbb{1}[i \in S_t] \cdot c_t(i,j) \cdot \mathbb{1}[j_t \leq j] | S_t] = q_{i,j}(S_t) \cdot \mathbb{P}[j_t \leq j | S_t]$$

By summation over time step  $1, 2, \dots, t$ , we have

$$\begin{aligned} \sum_{\tau=1}^t \mathbb{E}[\mathbb{1}[i \in S_\tau] \cdot c_\tau(i,j) \cdot \mathbb{1}[j_\tau \leq j] | S_\tau] &= \sum_{\tau=1}^t q_{i,j}(S_\tau) \cdot \mathbb{P}[j_\tau \leq j | S_\tau] \\ &= \sum_{\tau=1}^t \mathbb{E}[q_{i,j}(S_\tau) \cdot \mathbb{1}[j_\tau \leq j] | S_\tau], \end{aligned}$$

which implies that

$$\mathbb{E}\left[\sum_{\tau \leq t, i \in S_\tau, j_\tau \leq j} c_\tau(i,j) \middle| S_1, S_2, \dots, S_t\right] = \mathbb{E}\left[\sum_{\tau \leq t, j_\tau \leq j} q_{i,j}(S_\tau) \middle| S_1, \dots, S_t\right]$$

Notice that  $S_t$  is  $\mathcal{F}_{t-1}$ -measurable. By the definition of  $q_{i,j}(S_\tau)$ , we have

$$\mathbb{E}\left[\sum_{\tau \leq t, i \in S_\tau, j_\tau \leq j} c_\tau(i,j) - q_{i,j}(S_\tau) \middle| \mathcal{F}_{t-1}\right] = 0.$$

If we count the number of  $\tau$  that satisfies  $i \in S_\tau$  and  $j_\tau \leq j$  is exactly  $SC_t(i, j) = \sum_{\tau=1}^t \mathbb{1}[i \in S_\tau, j_\tau \leq j]$ . Therefore, by Azuma-Hoeffding inequality, we have for fixed  $SC_t(i, j)$ , with probability at least  $1 - \delta$ ,

$$\left| \sum_{\tau \leq t, i \in S_\tau, j_\tau \leq j} c_\tau(i, j) - \sum_{\tau \leq t, i \in S_\tau, j_\tau \leq j} q_{i,j}(S_\tau) \right| \leq \sqrt{2SC_t(i, j) \log(T/\delta)},$$

By union inequality, we have

$$\left| \sum_{\tau \leq t, i \in S_\tau, j_\tau \leq j} c_\tau(i, j) - \sum_{\tau \leq t, i \in S_\tau, j_\tau \leq j} q_{i,j}(S_\tau) \right| \leq \sqrt{8SC_t(i, j) \log(NMT)},$$

holds for any  $t \in [T]$ ,  $SC_t(i, j)$ , and  $(i, j) \in [N] \times [M]$  with probability at least  $1 - T^{-2}$ . We denote this good event as  $\mathcal{E}_1$  which satisfies  $\mathbb{P}[\neg \mathcal{E}_1] \leq T^{-2}$ .

We recall the definition of  $\hat{q}_{i,j}^t$  given in Algorithm 1

$$\hat{q}_{i,j}^t = \frac{C_t(i, j)}{SC_t(i, j)} = \frac{\sum_{\tau \leq t} c_\tau(i, j)}{SC_t(i, j)} = \frac{\sum_{\tau \leq t} \mathbb{1}[i \in S_\tau] \cdot c_\tau(i, j) \mathbb{1}[j_\tau \leq j]}{SC_t(i, j)}.$$

Under this good event  $\mathcal{E}_1$ , we have for every  $t \in [T]$  and  $(i, j) \in [N] \times [M]$ ,

$$\left| \hat{q}_{i,j}^t - \frac{\sum_{\tau \leq t, i \in S_\tau, j_\tau \leq j} q_{i,j}(S_\tau)}{SC_t(i, j)} \right| \leq \sqrt{8 \frac{\log(NMT)}{SC_t(i, j)}}$$

Below we bound the difference between  $q_{i,j}^*(S_t)$  and  $q_{i,j}(S_t)$  for any  $S_t \in \mathcal{S}$ . For given  $(i, j)$  with  $i \in S_t$ , we have

$$\begin{aligned} q_{i,j}^*(S_t) - q_{i,j}(S_t) &= \mathbb{P}[I_t \ni i, j_t = i \mid j_t \leq j, S_t] - \mathbb{P}[i_t = i, j_t = j \mid j_t \leq j, S_t] \\ &= \mathbb{P}[I_t \ni i, i_t \neq i, j_t = j \mid j_t \leq j, S_t] \\ &\leq \sum_{k \in S_t, k \neq i} \frac{\mathbb{P}[X_i \in M_j] \mathbb{P}[X_k \in M_j]}{\mathbb{P}[X_i \in M_{\leq j}] \mathbb{P}[X_k \in M_{\leq j}]} \\ &\leq (K-1) \cdot \frac{(L\epsilon)^2}{(j\epsilon/L)^2} = (K-1) \cdot L^4/j^2, \end{aligned}$$

where the last inequality holds by Assumption 4.1 and  $\mathbb{P}[X_i \in M_{\leq j}] = \sum_{j'=1}^j p_{i,j}^* \leq j \frac{\epsilon}{L}, \forall i \in [N]$ .

Notice that for every  $S_t \in \mathcal{S}$  and  $i \in S_t, j \in [M]$ , we have

$$\begin{aligned} q_{i,j}^*(S_t) &= \mathbb{P}[I_t \ni i, j = j_t \mid j_t \leq j, S_t] \\ &= \frac{\mathbb{P}[I_t \ni i, j = j_t \mid S_t]}{\mathbb{P}[j_t \leq j \mid S_t]} \\ &= \frac{\mathbb{P}[X_i(t) \in M_j \ \& \ X_k(t) \in M_{\leq j}, \forall k \in S_t \mid S_t]}{\mathbb{P}[X_k(t) \in M_{\leq j}, \forall k \in S_t \mid S_t]} \\ &= \frac{\mathbb{P}[X_i \in M_j] \cdot \mathbb{P}[X_k \in M_{\leq j}, \forall k \in S, k \neq i]}{\mathbb{P}[X_i \in M_{\leq j}] \cdot \mathbb{P}[X_k \in M_{\leq j}, \forall k \in S, k \neq i]} \\ &= \frac{\mathbb{P}[X_i \in M_j]}{\mathbb{P}[X_i \in M_{\leq j}]} \\ &= \mathbb{P}[X_i \in M_j \mid X_i \in M_{\leq j}] \\ &= q_{i,j}^*. \end{aligned}$$

Therefore, we have

$$|\hat{q}_{i,j}^t - q_{i,j}^*| = \left| \hat{q}_{i,j}^t - \frac{\sum_{\tau \leq t, i \in S_\tau, j_\tau \leq j} q_{i,j}^*(S_\tau)}{SC_t(i,j)} \right| \leq \sqrt{8 \frac{\log(NMt)}{SC_t(i,j)}} + (K-1) \cdot (L^4/j^2)$$

□

#### A.4 Optimistic Estimation

**Lemma A.4.** For  $\beta_{i,j}^t$  given in Eq. (6), under event  $\mathcal{E}_0$  and  $\mathcal{E}_1$ , we have

$$\bar{q}_{i,j}^t \geq q_{i,j}^*.$$

Moreover, by the offline  $(1-\epsilon)$ -approximated optimization oracle PTAS (Chen et al., 2013), we have

$$\bar{r}_q(S_t; \bar{\mathbf{q}}^t) \geq (1-\epsilon) \cdot \bar{r}_q(S^*; \bar{\mathbf{q}}^t).$$

*Proof.* Notice that in Algorithm 1 we define

$$\bar{q}_{i,j}^t = \min \left\{ \hat{q}_{i,j}^t + \beta_{i,j}^t + \frac{(K-1)L^4}{j^2}, 1 \right\}.$$

By Lemma A.3, we have under  $\neg\mathcal{E}_0$  and  $\mathcal{E}_1$ ,

$$\hat{q}_{i,j}^t \geq q_{i,j}^* - \beta_{i,j}^t - \frac{(K-1)L^4}{j^2},$$

where the inequality holds by the definition of  $\beta_{i,j}^t$  in Eq. (6) and  $SC_{t-1}(i,j) \leq SC_t(i,j)$ . Since  $q_{i,j}^* \leq 1$ , we have

$$\bar{q}_{i,j}^t \geq q_{i,j}^*.$$

Since in Algorithm 1, we set action  $S_t \leftarrow \text{PTAS}(\hat{\mathbf{p}}^t)$  where  $\hat{\mathbf{p}}^t$  is converted from  $\hat{\mathbf{q}}^t$  by Eq. (3). Then by Lemmas 4.3 and 4.4, we have

$$\bar{r}_q(S_t; \bar{\mathbf{q}}^t) = \bar{r}(S_t; \bar{\mathbf{p}}^t) \geq (1-\epsilon) \max_{S \in \mathcal{S}} \bar{r}(S; \bar{\mathbf{p}}^t) \geq (1-\epsilon) \bar{r}(S^*; \bar{\mathbf{p}}^t) = (1-\epsilon) \bar{r}_q(S^*; \bar{\mathbf{q}}^t).$$

□

#### A.5 Regret Decomposition

**Lemma A.5.** Denote  $Q_j^*(S_t) := \prod_{k \in S_t, j' > j} (1 - q_{k,j'}^*)$ . We have

$$|\bar{r}_q(S_t; \bar{\mathbf{q}}^t) - \bar{r}_q(S_t; \mathbf{q}^*)| \leq 2 \sum_{i \in S_t, j \in [M]} Q_j^*(S_t) \cdot v_j \cdot |\bar{q}_{i,j}^t - q_{i,j}^*|. \quad (19)$$

*Proof.* This lemma is given by directly apply Lemma 3.3 in Wang et al. (2023) by definition of  $\bar{r}_q$  in Eq. (17). □

**Lemma A.6.** Under Assumption 4.1, we can bound the regret of Algorithm 1 by

$$\mathcal{R}(T) \leq \mathbb{E} \left[ \sum_{t=1}^T \text{Bonus}_t + \text{Bias}_t \middle| \mathcal{E}_0, \mathcal{E}_1 \right] + 3T\epsilon + T^{-1},$$

where  $\text{Bonus}_t$  and  $\text{Bias}_t$  is defined by

$$\text{Bonus}_t := 4 \sum_{i \in S_t, j \in [M]} Q_j^*(S_t) \cdot v_j \cdot \beta_{i,j}^t, \quad (20)$$

and

$$\text{Bias}_t := 4 \sum_{i \in S_t, j \in [M]} Q_j^*(S_t) \cdot v_j \cdot (K-1) \frac{L^4}{j^2}. \quad (21)$$

*Proof.* This lemma formalize the first three steps of proof sketch. Denote  $\Delta_t := r^*(S^*) - r^*(S_t)$ , we have

$$\mathcal{R}(T) = \mathbb{E} [\Delta_t].$$

By Lemma 4.2, we have

$$\Delta_t \leq \bar{r}(S^*; \mathbf{p}^*) - \bar{r}(S_t; \mathbf{p}^*) + 2\epsilon.$$

Then we have

$$\begin{aligned} \mathcal{R}(T) &\leq \mathbb{P}[\mathcal{E}_0] \cdot \mathbb{E} \left[ \sum_{t=1}^T \Delta_t \middle| \mathcal{E}_0 \right] + \mathbb{P}[\neg \mathcal{E}_0] \cdot T \\ &\leq \mathbb{E} \left[ \sum_{t=1}^T \Delta_t \middle| \mathcal{E}_0 \right] \\ &\leq \mathbb{E} \left[ \sum_{t=1}^T \bar{r}(S^*; \mathbf{p}^*) - \bar{r}(S_t; \mathbf{p}^*) \middle| \mathcal{E}_0 \right] + 2T\epsilon, \end{aligned}$$

where the first inequality holds by property of conditional expectations and  $\Delta_t \leq 1$  and the second inequality is due to  $\mathbb{P}[\neg \mathcal{E}_0] = 0$ .

Notice that under  $\mathcal{E}_0$  and  $\mathcal{E}_1$ , by Lemmas 4.4 and A.4, we have

$$\bar{r}_q(S_t; \mathbf{q}^t) \geq (1-\epsilon)\bar{r}_q(S^*; \bar{\mathbf{q}}_t) \geq (1-\epsilon)\bar{r}_q(S^*; \mathbf{q}^*).$$

Then with Lemma A.2, we have

$$\begin{aligned} \mathcal{R}(T) &\leq \mathbb{E} \left[ \sum_{t=1}^T \bar{r}_q(S^*; \mathbf{q}^*) - \bar{r}_q(S_t; \mathbf{q}^*) \middle| \mathcal{E}_0 \right] + 2T\epsilon \\ &\leq \mathbb{E} \left[ \sum_{t=1}^T \bar{r}_q(S^*; \mathbf{q}^*) - \bar{r}_q(S_t; \mathbf{q}^*) \middle| \mathcal{E}_0, \mathcal{E}_1 \right] + \mathbb{P}[\neg \mathcal{E}_1] \cdot T + 2T\epsilon \\ &\leq \mathbb{E} [\bar{r}_q(S_t; \mathbf{q}^t) - \bar{r}_q(S_t; \mathbf{q}^*)] + 3T\epsilon + T^{-1}, \end{aligned}$$

where the last inequality holds by  $\epsilon \bar{r}_q(S^*; \mathbf{p}^*) \leq \epsilon$  and  $\mathbb{P}[\neg \mathcal{E}_1] \leq T^{-2}$  shown in Lemma A.3.

Therefore, applying Lemma 4.7, we get

$$\mathcal{R}(T) \leq \mathbb{E} \left[ \sum_{t=1}^T \text{Bonus}_t + \text{Bias}_t \middle| \mathcal{E}_0, \mathcal{E}_1 \right] + 3T\epsilon + T^{-1},$$

where  $\text{Bonus}_t$  and  $\text{Bias}_t$  is defined in Eqs. (20) and (21).  $\square$

## A.6 Bounding the Bonus Terms

We apply similar methods in Wang and Chen (2017); Liu et al. (2023b) to give the bounds of  $\sum_t \text{Bonus}_t$ . We first give the following definitions.

**Definition A.7** (Wang and Chen (2017, Definition 5)). Let  $(i, j) \in [N] \times [M]$  be the index of binary arm and  $l$  be a positive natural number, define the triggering probability group (of actions)

$$S_j^l = \{S \in \mathcal{S} \mid 2^{-l} < Q_j^*(S) \leq 2^{-l+1}\}.$$

Notice  $\{S_j^l\}_{l \geq 1}$  forms a partition of  $\{S \in \mathcal{S} \mid Q_j^*(S) > 0\}$ .

**Definition A.8** (Wang and Chen (2017, Definition 6)). For each group  $S_j^l$  (Definition A.7), we define a corresponding counter  $N_{i,j}^l$ . In a run of a learning algorithm, the counters are maintained in the following manner. All the counters are initialized to 0. In each round  $t$ , if the action  $S_t$  is chosen, then update  $N^l(i, j)$  to  $N^l(i, j) + 1$  for every  $(i, j)$  that  $i \in S_t, S_t \in S_j^l$ . Denote  $N_t^l(i, j)$  at the end of round  $t$  with  $N^l(i, j)$ . In other words, we can define the counters with the recursive equation below:

$$N_t^l(i, j) = \begin{cases} 0, & \text{if } t = 0, \\ N_{t-1}^l(i, j) + 1, & \text{if } t > 0, i \in S_t, S_t \in S_j^l, \\ N_{t-1}^l(i, j), & \text{otherwise.} \end{cases}$$

**Definition A.9** (Wang and Chen (2017, Definition 7)). Given a series of integers  $\{l_{i,j}^{\max}\}_{i \in [N], j \in [M]}$ , we say that the triggering is nice at the beginning of round  $t$  (with respect to  $l_{i,j}^{\max}$ ), if for every group  $S_j^l$  (Definition A.7) identified by binary arm  $(i, j)$  and  $1 \leq l \leq l_{i,j}^{\max}$ , as long as

$$\sqrt{\frac{8 \log(NMT)}{\frac{1}{3} N_{t-1}^l(i, j) \cdot 2^{-l}}} \leq 1,$$

there is  $SC_{t-1}(i, j) \geq \frac{1}{3} N_{t-1}^l(i, j) \cdot 2^{-l}$ . We denote this event with  $\mathcal{E}_2(t)$ . It implies

$$\beta_{i,j}^t = \sqrt{\frac{8 \log(NMT)}{SC_{t-1}(i, j)}} \leq \sqrt{\frac{8 \log(NMT)}{\frac{1}{3} N_{t-1}^l(i, j) \cdot 2^{-l}}}.$$

Therefore, we show that  $\mathcal{E}_2(t)$  happens with high probability for every  $t$ .

**Lemma A.10** (Wang and Chen (2017, Lemma 4)). *For a series of integers  $\{l_{i,j}^{\max}\}_{i \in [N], j \in [M]}$ ,*

$$\mathbb{P}[\neg \mathcal{E}_2(t)] \leq \sum_{i \in [N], j \in [M]} l_{i,j}^{\max} t^{-2},$$

for every round  $t \geq 1$ .

*Proof.* We prove this lemma by showing  $\mathbb{P}[N_{t-1}^l(i, j) = s, SC_{t-1}(i, j) \leq \frac{1}{3} N_{t-1}^l(i, j) \cdot 2^{-l}] \leq t^{-3}$ , for any fixed  $s$  with  $0 \leq s \leq t - 1$  and  $\sqrt{\frac{8 \log(NMT)}{\frac{1}{3} s \cdot 2^{-l}}} \leq 1$ . Let  $t_k$  be the round that  $N^l(i, j)$  is increased for the  $k$ -th time, for  $1 \leq k \leq s$ . Let  $Z_k = \mathbb{1}[S_{t_k} \ni i, j_{t_k} \leq j]$  be a Bernoulli variable, that is,  $SC_{t_k}(i, j)$  increase in round  $t_k$ . When fixing the action  $S_{t_k}$ ,  $Z_k$  is independent

from  $Z_1, \dots, Z_{k-1}$ . Since  $S_{t_k} \in S_j^l$ ,  $\mathbb{E}[Z_k \mid Z_1, \dots, Z_{k-1}] \geq 2^{-l}$ . Let  $Z = Z_1 + \dots + Z_s$ . By multiplicative Chernoff bound ([Upfal and Mitzenmacher, 2005](#)), we have

$$\mathbb{P}\left\{Z \leq \frac{1}{3}s \cdot 2^{-l}\right\} \leq \exp\left(-\frac{\left(\frac{2}{3}\right)^2 s \cdot 2^{-l}}{2}\right) \leq \exp\left(-\frac{\left(\frac{2}{3}\right)^2 18 \log t}{2}\right) < \exp(-3 \log t) = t^{-3}.$$

By the definition of  $SC_{t-1}(i, j)$  and the condition  $N_{t-1}^l(i, j) = s$ , we have  $SC_{t-1}(i, j) \geq Z$ . Thus

$$\begin{aligned}\mathbb{P}[N_{t-1}^l(i, j) = s, SC_{t-1}(i, j) \leq \frac{1}{3}N_{t-1}^l(i, j) \cdot 2^{-l}] \\ \leq \mathbb{P}[N_{t-1}^l(i, j) = s, Z \leq \frac{1}{3}s \cdot 2^{-l}] \\ \leq \mathbb{P}[Z \leq \frac{1}{3}s \cdot 2^{-l}] \leq t^{-3}.\end{aligned}$$

By taking  $i, j$  over  $[N] \times [M]$ ,  $l$  over  $1, \dots, l_{i,j}^{\max}$ ,  $s$  over  $0, \dots, t-1$  and applying the union bound, the lemma holds.  $\square$

**Lemma A.11.** *For given constant  $C$ , we have*

$$\sum_{t=1}^T \text{Bonus}(t) \leq 16NM + 12288 \frac{KNM^2 \log(NMT)}{C} + TC + \frac{\pi^2}{6} \left\lceil \log_2 \frac{16KM}{C} \right\rceil.$$

*Proof.* For given constant  $C$ , we can define the following notations.

$$l_{i,j}^{\max} := \left\lceil \log_2 \frac{16KM}{C} \right\rceil, \quad \forall (i, j) \in [N] \times [M], \quad (22)$$

and for every integer  $l$ ,

$$\kappa_{l,T}(C, s) := \begin{cases} 2 \cdot 2^{-l} & s = 0 \\ \sqrt{96 \cdot 2^{-l} \log(NMT)/s} & 1 \leq s \leq B_{l,T}(C), \\ 0 & s > B_{l,T}(C) \end{cases} \quad (23)$$

where  $B_{l,T}(C)$  is given by

$$B_{l,T}(C) := \left\lfloor 6144 \cdot 2^{-l} K^2 M^2 \log(NMT)/C^2 \right\rfloor. \quad (24)$$

By [Wang and Chen \(2017, Lemma 5\)](#), if  $\text{Bonus}(t) \geq C$ , under event  $\mathcal{E}_2(t)$ , we have

$$\text{Bonus}(t) \leq \sum_{i \in S_t, j \in [M]} \kappa_{l_{i,j}, T}(C, N_{t-1}^{l_i}(i, j)),$$

where  $l_{i,j}$  is the index of group  $S_j^{l_{i,j}} \ni S_t$ . This is because we have

$$\begin{aligned}\text{Bonus}(t) &\leq -C + 8 \sum_{i \in S_t, j \in [M]} Q_j^*(S_t) \cdot (j-1)\epsilon \cdot \min\{\beta_{i,j}^t, 1\} \\ &\leq 8 \sum_{i \in S_t, j \in [M]} \left( Q_j^*(S_t) \cdot \min\{\beta_{i,j}^t, 1\} - \frac{C}{8KM} \right)\end{aligned}$$

**Case 1:**  $1 \leq l_{i,j} \leq l_{i,j}^{\max}$ . We have

$$Q_j^*(S_t) \leq 2 \cdot 2^{-l_{i,j}}.$$

Under  $\mathcal{E}_2(t)$ , we have

$$\min \{\beta_{i,j}^t, 1\} = \min \left\{ \sqrt{\frac{8 \log(NMT)}{SC_{t-1}(i,j)}}, 1 \right\} \leq \min \left\{ \sqrt{\frac{8 \log(NMT)}{\frac{1}{3}N_{t-1}^{l_{i,j}}(i,j) \cdot 2^{-l_{i,j}}}}, 1 \right\},$$

and

$$\begin{aligned} Q_j^*(S_t) \cdot \min \{\beta_{i,j}^t, 1\} &\leq 2 \cdot 2^{-l_{i,j}} \cdot \min \left\{ \sqrt{\frac{8 \log(NMT)}{\frac{1}{3}N_{t-1}^{l_{i,j}}(i,j) \cdot 2^{-l_{i,j}}}}, 1 \right\} \\ &\leq \min \left\{ \sqrt{\frac{96 \cdot 2^{-l_{i,j}} \log(NMT)}{N_{t-1}^{l_{i,j}}(i,j)}}, 2 \cdot 2^{-l_{i,j}} \right\}. \end{aligned} \quad (25)$$

If  $N_{t-1}^{l_{i,j}}(i,j) \geq B_{l_{i,j},T}(C) + 1$ , then

$$\sqrt{\frac{96 \cdot 2^{-l_{i,j}} \log(NMT)}{N_{t-1}^{l_{i,j}}(i,j)}} \leq \frac{C}{8KM},$$

which implies  $Q_j^*(S_t) \cdot \min \{\beta_{i,j}^t, 1\} - C/8KM \leq 0$ .

If  $N_{t-1}^{l_{i,j}}(i,j) = 0$ , we have  $Q_j^*(S_t) \cdot \min \{\beta_{i,j}^t, 1\} \leq Q_j^*(S_t) \leq 2 \cdot 2^{-l_{i,j}}$ , which implies

$$Q_j^*(S_t) \cdot \min \{\beta_{i,j}^t, 1\} - \frac{C}{8KM} \leq \kappa_{l_{i,j},T}(C, 0)$$

Otherwise, for  $1 \leq N_{t-1}^{l_{i,j}}(i,j) \leq B_{l_{i,j},T}(C)$ , we have  $Q_j^*(S_t) \cdot \min \{\beta_{i,j}^t, 1\} \leq \kappa_{l_{i,j},T}(C, N_{t-1}^{l_{i,j}}(i,j))$  by Eqs. (23) and (25). Therefore, we get

$$Q_j^*(S_t) \cdot \min \{\beta_{i,j}^t, 1\} - \frac{C}{8KM} \leq \kappa_{l_{i,j},T}(C, N_{t-1}^{l_{i,j}}(i,j))$$

**Case 2:**  $l_{i,j} \geq l_{i,j}^{\max} + 1$ . We have

$$Q_j^*(S_t) \cdot \min \{\beta_{i,j}^t, 1\} \leq 2 \cdot 2^{-l_{i,j}} \leq 2 \cdot \frac{C}{16KM} \leq \frac{C}{8KM},$$

which shows that  $Q_j^*(S_t) \cdot \min \{\beta_{i,j}^t, 1\} - C/8KM \leq 0$ . If  $N_{t-1}^{l_{i,j}}(i,j) = 0$ . Therefore, we finally get

$$\text{Bonus}(t) \leq 8 \sum_{i \in S_t, j \in [M]} \kappa_{l_{i,j},T}(C, N_{t-1}^{l_{i,j}}(i,j)),$$

for the case of good event  $\mathcal{E}_2(t)$  happens and  $\text{Bonus}(t) \geq C$ .

Notice that under good events  $\mathcal{E}_0, \mathcal{E}_1$ , we have

$$\begin{aligned} \sum_{t=1}^T \text{Bonus}(t) &\leq \sum_{t=1}^T \mathbb{1}[\{\text{Bonus}(t) \geq C\} \cap \mathcal{E}_2(t)] \cdot \text{Bonus}(t) + T \cdot C + \sum_{t=1}^T \mathbb{P}[\mathcal{E}_2(t)] \\ &\leq \underbrace{\sum_{t=1}^T 8 \cdot \sum_{i \in S_t, j \in [M]} \kappa_{l_{i,j},T}(C, N_{t-1}^{l_{i,j}}(i,j))}_{(I)} + TC + \frac{\pi^2}{6} \cdot \max_{i \in [N], j \in [M]} l_{i,j}^{\max}. \end{aligned}$$

where the first inequality is due to  $\text{Bonus}(t) \leq 1$  and definition, and the second one is due to Lemma A.10. The key is bounding  $(I)$ :

$$\begin{aligned} (I) &= 8 \cdot \sum_{i \in [N], j \in [M]} \sum_{l=1}^{\infty} \sum_{s=0}^{N_{T-1}^l(i,j)} \kappa_l(C, s) \\ &= 8 \cdot \sum_{i \in [N], j \in [M]} \sum_{l=1}^{\infty} \left( 2 \cdot 2^{-l} + \sum_{s=1}^{B_{l,T}(C)} \sqrt{\frac{96 \cdot 2^{-l} \log(NMT)}{s}} \right) \\ &\leq 8 \cdot \sum_{i \in [N], j \in [M]} \sum_{l=1}^{\infty} \left( 2 \cdot 2^{-l} + 2 \cdot \sqrt{96 \cdot 2^{-l} \log(NMT)} \cdot \sqrt{B_{l,T}(C)} \right), \end{aligned}$$

where the inequality holds by the fact that  $\sum_{s=1}^n \sqrt{1/s} \leq 2\sqrt{n}$ . Therefore, by the definition of  $B_{l,T}(C)$  in Eq. (24), we have

$$\begin{aligned} (I) &\leq 8 \cdot \sum_{i \in [N], j \in [M]} \sum_{l=1}^{\infty} \left( 2 \cdot 2^{-l} + 1536 \cdot \frac{2^{-l} KM \log(NMT)}{C} \right) \\ &= 8 \cdot \sum_{i \in [N], j \in [M]} \left( 2 + 1536 \cdot \frac{KM \log(NMT)}{C} \right) \cdot \left( \sum_{l=1}^{\infty} 2^{-l} \right) \\ &\leq 16NM + 12288 \frac{KNM^2 \log(NMT)}{C}. \end{aligned}$$

Therefore, we get

$$\sum_{t=1}^T \text{Bonus}(t) \leq 16NM + 12288 \frac{KNM^2 \log(NMT)}{C} + TC + \frac{\pi^2}{6} \left\lceil \log_2 \frac{16KM}{C} \right\rceil.$$

□

## A.7 Bounding the Bias Terms

**Lemma A.12.** *Under Assumption 4.1, we have*

$$\sum_{t=1}^T \text{Bias}(t) \leq 4K^2 L^4 T \epsilon \log(M+1).$$

*Proof.* Notice that  $\sum_{j \in [M]} 1/j \leq \log(M+1)$  for  $\epsilon < 1/2$ , we have

$$\begin{aligned} \text{Bias}(t) &\leq 4K \cdot \sum_{i \in S_t, j \in [M]} Q_j^*(S_t) \cdot \epsilon L^4 / j \\ &= 4K^2 L^4 \epsilon \cdot \sum_{j \in [M]} \frac{1}{j} \\ &\leq 4K^2 L^4 \epsilon \log(M+1). \end{aligned}$$

Therefore, we have

$$\sum_{t=1}^T \text{Gap}(t) \leq 4K^2 L^4 T \epsilon \log(M+1).$$

□

## A.8 Proof of Theorem 4.6

**Theorem A.13** (Formal version of Theorem 4.6). *By setting  $\beta_{i,j}^t$  in Eq. (6) and  $\epsilon < 1/2$ , we can control the regret of Algorithm 1 under Assumption 4.1 by*

$$\begin{aligned}\mathcal{R}(T) &\leq 12289\sqrt{NKM^2T\log(NMT)} + T\epsilon(4KL^4\log(M+1) + 3) \\ &\quad + 16NM + \pi^2\left(\log_2(\sqrt{KM^2T\log(NMT)/N}) + 5\right)/6 + T^{-1} \\ &= \tilde{O}\left(\sqrt{NKM^2T} + L^4K^2T\epsilon\right),\end{aligned}$$

where  $M = \lceil 1/\epsilon \rceil$ . If we further take  $\epsilon = O\left(L^{-2}K^{-\frac{3}{4}}N^{\frac{1}{4}}T^{-\frac{1}{4}}\right)$ , we have

$$\mathcal{R}(T) = \tilde{O}(L^2N^{\frac{1}{4}}K^{\frac{5}{4}}T^{\frac{3}{4}}).$$

*Proof.* By Lemma A.6, we have

$$\mathcal{R}(T) \leq \mathbb{E}\left[\sum_{t=1}^T \text{Bonus}_t + \text{Bias}_t \middle| \mathcal{E}_0, \mathcal{E}_1\right] + 3T\epsilon + T^{-1},$$

Take constant  $C$  as

$$C := \sqrt{\frac{NKM^2\log(NMT)}{T}}. \quad (26)$$

Then Lemma A.11 shows that

$$\sum_{t=1}^T \text{Bonus}(t) \leq 16NM + 12289\sqrt{NM^2KT\log(NMT)} + \pi^2\left(\log_2(\sqrt{KMT\log(NMT)/N}) + 5\right)/6.$$

Lemma A.12 demonstrates that

$$\sum_{t=1}^T \text{Bias}(t) \leq 4K^2L^4T\epsilon\log(M+1).$$

Therefore, by calculating the summation of the bonus and bias terms, we can bound the regret by

$$\begin{aligned}\mathcal{R}(T) &\leq \mathbb{E}\left[\sum_{t=1}^T \text{Bonus}(t) + \text{Gap}(t) \middle| \mathcal{E}_0, \mathcal{E}_1\right] + T^{-1} + 3T\epsilon \\ &\leq 12289\sqrt{NKM^2T\log(NMT)} + T\epsilon(4KL^4\log(M+1) + 3) \\ &\quad + 16NM + \pi^2\left(\log_2(\sqrt{KM^2T\log(NMT)/N}) + 5\right)/6 + T^{-1} \\ &= \tilde{O}\left(\sqrt{NKM^2T} + L^4K^2T\epsilon\right),\end{aligned}$$

which finishes the proof.  $\square$

## B Omitted proofs in Section 5

This proof mainly applies the techniques for general linear bandits Liu et al. (2024a); Lee et al. (2024). Given action  $S \in \mathcal{S}$  to the environment, we assume that  $\ell_S$  is the random variable of the loss, i.e.,  $\mathbb{E}[\ell_S] = \ell^*(S)$ .

We have

$$\begin{aligned} g_t(\theta; \lambda) &:= -\nabla_\theta \mathcal{L}_t(\theta; \lambda) + \sum_{i < t} \ell_i \psi(S_i) \\ &= \sum_{i < t} \frac{1}{\psi(S_i)^\top \theta} \cdot \psi(S_i) - \lambda \theta. \\ H_t(\theta; \lambda) &:= \nabla_\theta^2 \mathcal{L}(\theta; \mathcal{H}_t) \\ &= -\nabla_\theta g_t(\theta; \lambda) \\ &= \lambda I + \sum_{i < t} \frac{\psi(S_i) \psi(S_i)^\top}{(\psi(S_i)^\top \theta)^2}. \end{aligned}$$

### B.1 Concentration Argument for MLE

**Lemma B.1** (MLE Concentration). *For  $L^* := \sup_{S \in \mathcal{S}} \ell^*(S)$ ,  $M_1 := L^*/\sqrt{2}$ , and  $V = \sup\{\|\theta\|_2 : \theta \in \Theta\}$ , set*

$$\lambda_t := \max \left\{ 1, \frac{2dM_1}{V} \cdot \log \left( e \sqrt{1 + \frac{tL^*}{d}} + \frac{1}{\delta} \right) \right\}, \quad (27)$$

and

$$\gamma_t(\delta, \lambda_t) := \sqrt{\lambda_t} \left( \frac{1}{2M_1} + V \right) + \frac{2M_1 d}{\sqrt{\lambda_t}} \left( \log(2) + \frac{1}{2} \log \left( 1 + \frac{tL^*}{\lambda_t d} \right) \right) + \frac{2M_1}{\sqrt{\lambda_t}} \log(1/\delta). \quad (28)$$

Then we have with probability at least  $1 - \delta$ ,

$$\theta^* \in C_t(\hat{\theta}_t; \delta, \lambda_t) := \left\{ \theta \in \Theta : \left\| g_t(\theta; \lambda_t) - g_t(\hat{\theta}_t; \lambda_t) \right\|_{H_t^{-1}(\theta; \lambda_t)} \leq \gamma_t(\delta, \lambda_t) \right\}, \quad (29)$$

holds for any  $t \in [T]$ . We denote the confidence set as  $C_t(\hat{\theta}_t; \delta, \lambda_t)$  and this good event as  $\Xi$ .

*Proof.* For simplicity, we denote the filtration of history as  $\mathcal{H}_t := (S_1, Y_1, \dots, S_{t-1}, Y_{t-1}, S_t)$ . Then we have

$$\ell_t \sim \exp(\psi(S_t)^\top \theta^*), \quad \mathbb{E}[\ell_t | \mathcal{H}_t] = \frac{1}{\psi(S_t)^\top \theta^*},$$

by the property of exponential distribution and definition of  $\psi(S)$ . Since we have

$$\hat{\theta}_t \leftarrow \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \mathcal{L}_t(\theta; \lambda_t),$$

by Algorithm 2. Then by KKT condition, we have

$$\left. \frac{\partial \mathcal{L}_t(\theta; \lambda_t)}{\partial \theta} \right|_{\theta=\hat{\theta}_t} = 0 \Rightarrow g_t(\hat{\theta}_t; \lambda_t) - \sum_{i < t} \ell_i \psi(S_i) = 0$$

Notice that by definition of  $g_t$ ,

$$g_t(\theta^*; \lambda_t) = \sum_{i < t} \frac{1}{\psi(S_i)^\top \theta^*} \cdot \psi(S_i) - \lambda_t \theta^*.$$

Denote  $\varepsilon_t := \ell_t - \mathbb{E}[\ell_t | \mathcal{H}_t] = \ell_t - 1/(\psi_t(S_t)^\top \theta^*)$ , we have

$$g_t(\hat{\theta}_t; \lambda_t) - g_t(\theta^*; \lambda_t) = \sum_{i < t} \varepsilon_i \psi(S_i) + \lambda_t \theta^*.$$

Fix  $s \geq 0$ , we have

$$\begin{aligned} \mathbb{E}[\exp(s\varepsilon_t) | \mathcal{H}_t] &= \mathbb{E}\left[\exp\left(s\ell_t - \frac{s}{\psi_t(S_t)^\top \theta^*}\right)\right] \\ &= \exp\left(-\frac{s}{\psi_t(S_t)^\top \theta^*}\right) \cdot \mathbb{E}[\exp(s\ell_t) | \mathcal{H}_t], \end{aligned}$$

and by calculation,

$$\begin{aligned} \mathbb{E}[\exp(s\varepsilon_t) | \mathcal{H}_{t-1}] &= \exp\left(-\frac{s}{\psi(S_t)^\top \theta^*}\right) \cdot \mathbb{E}[\exp(s\ell_t) | \mathcal{H}_t] \\ &= \exp\left(-\frac{s}{\psi(S_t)^\top \theta^*}\right) \cdot \int_{(0,+\infty)} \psi(S_t)^\top \theta^* \exp(-(\psi(S_t)^\top \theta^* - s)y) dy \\ &= \exp\left(-\frac{1}{\ell_t^*} s + \log(\ell_t^*) - \log(\ell_t^* - s)\right), \end{aligned}$$

where we use  $\ell_t^* := \psi_t(S_t)^\top \theta^*$  for simplicity. Consider the case for  $s < \ell_t^*$ , by intermediate value theorem, we have

$$\log(\ell_t^*) - \log(\ell_t^* - s) = s \cdot \frac{1}{\ell_t^*} - \frac{s^2}{2\xi^2},$$

for some  $\xi \in [\ell_t^* - s, \ell_t^*]$ . We further denote

$$L^* = \sup_{S \in \mathcal{S}} \ell^*(S). \quad (30)$$

Therefore, we can set constant  $0 \leq s \leq L^*$ , which gives

$$\log(\ell_t^*) - \log(\ell_t^* - s) \leq s \cdot \frac{1}{\ell_t^*} - \frac{s^2}{(\ell_t^*)^2}.$$

Denote  $\nu_{t-1} := -1/(\psi(S_t)^\top \theta^*)^2$ . We have for some constant  $M_1 \geq L^*/\sqrt{2}$ , and  $|s| \leq 1/M_1$ ,

$$\mathbb{E}[\exp(s\varepsilon_t) | \mathcal{H}_t] \leq \exp(s^2 \nu_{t-1}).$$

Applying [Janz et al. \(2024, Theorem 2\)](#) with  $S_t := \sum_{i < s} \varepsilon_i \psi(S_i)$ , we can show that with probability at least  $1 - \delta$ ,

$$\begin{aligned} \|g_t(\hat{\theta}_t; \lambda_t) - g_t(\theta^*; \lambda_t)\|_{H_t^{-1}(\theta^*; \lambda_t)} &\leq \left\| \sum_{i < t} \varepsilon_i \psi_i(S_i) \right\|_{H_t^{-1}(\theta^*; \lambda_t)} + \lambda_t \|\theta^*\|_{H_t^{-1}(\theta^*; \lambda_t)} \\ &\leq \frac{\sqrt{\lambda_t}}{2M_1} + \frac{2M_1}{\sqrt{\lambda_t}} \log\left(\frac{\det(H_t(\theta^*)^{1/2}/\lambda_t^{d/2})}{\delta}\right) + \frac{2M_1}{\sqrt{\lambda_t}} d \log(2) + \sqrt{\lambda_t} V, \end{aligned}$$

where  $V = \sup\{\|\theta\|_2 : \theta \in \Theta\}$ . Moreover, by definition of  $H_t(\theta^*; \lambda_t)$ , we have

$$\det(H_t(\theta^*; \lambda_t))/\lambda_t^d \leq \left(1 + \frac{tL^*}{\lambda_t d}\right)^d,$$

Therefore, for

$$\gamma_t(\delta, \lambda_t) \geq \sqrt{\lambda_t} \left( \frac{1}{2M_1} + V \right) + \frac{2M_1 d}{\sqrt{\lambda_t}} \left( \log(2) + \frac{1}{2} \log \left( 1 + \frac{tL^*}{\lambda_t d} \right) \right) + \frac{2M_1}{\sqrt{\lambda_t}} \log(1/\delta),$$

we have with probability at least  $1 - \delta$ ,

$$\theta^* \in C_t(\hat{\theta}_t; \delta, \lambda_t) := \left\{ \theta \in \Theta : \left\| g_t(\theta; \lambda_t) - g_t(\hat{\theta}_t; \lambda_t) \right\|_{H_t^{-1}(\theta; \lambda_t)} \leq \gamma_t(\delta, \lambda_t) \right\},$$

holds for any  $t \in [T]$ .  $\square$

## B.2 Proof of Theorem 5.1

**Theorem B.2** (Formal version of Theorem 5.1). *By setting  $\delta = 1/T$ ,  $\gamma_t(\delta)$  according to Eq. (28), and  $\lambda_t$  according to Eq. (27), Algorithm 2 enjoys the following regret guarantee:*

$$\begin{aligned} \mathcal{R}(T) &\leq 16\gamma \cdot \sqrt{dT} \cdot \sqrt{(\ell^*(S^*))^2(1 + L^*/\lambda) \cdot \log(1 + L^*T/d\lambda)} \\ &\quad + 256\gamma^2 \cdot dL^* \cdot \log(1 + L^*T/d\lambda) \cdot \left( \frac{\sup_{S \in \mathcal{S}} (\psi(S)^\top \theta^*)}{\ell^*(S^*)^3} + 2 \right) + 1, \end{aligned}$$

where  $\gamma := \sup_t \gamma_t(\delta)$  and  $\lambda_t := \inf_t \lambda_t$ .

*Proof.* Since we have  $X_i \sim \exp(\phi(i)^\top \theta^*)$ , then

$$\min_{i \in S} X_i \sim \exp \left( \sum_{i \in S} \phi(i)^\top \theta^* \right) = \exp(\psi(S)^\top \theta^*),$$

which shows that

$$\ell^*(S) = \mathbb{E} \left[ \min_{i \in S} X_i \right] = \frac{1}{\psi(S)^\top \theta^*}.$$

Therefore, by second-order Taylor expansion, we have for some  $\xi \in [\ell^*(S_t), \sup_t \ell^*(S_t)]$ ,

$$\begin{aligned} \mathcal{R}(T) &= \mathbb{E} \left[ \sum_{t=1}^T \ell^*(S_t) - \ell^*(S^*) \right] \\ &\leq \mathbb{P}[\Xi] \cdot \mathbb{E} \left[ \sum_{t=1}^T \frac{1}{\psi(S_t)^\top \theta^*} - \frac{1}{\psi(S^*)^\top \theta^*} \middle| \Xi \right] + \mathbb{P}[\neg \Xi] \cdot T \\ &\leq \mathbb{E} \left[ \underbrace{\sum_{t=1}^T \frac{1}{(\psi(S_t)^\top \theta^*)^2} \cdot (\psi(S^*)^\top \theta^* - \psi(S_t)^\top \theta^*)}_{\mathcal{R}_1(T)} \middle| \Xi \right] + \mathbb{E} \left[ \underbrace{\sum_{t=1}^T \frac{2}{\xi^3} \cdot (\psi(S^*)^\top \theta^* - \psi(S_t)^\top \theta^*)^2}_{\mathcal{R}_2(T)} \middle| \Xi \right] + 1. \end{aligned}$$

Under  $\Xi$ , we have  $\theta^* \in C_t(\hat{\theta}_t; \delta, \lambda_t)$  for every  $t \in [T]$ . Therefore, by Algorithm 2, we have

$$\psi(S^*)^\top \theta^* \leq \psi(S_t)^\top \tilde{\theta}_t. \quad (31)$$

Under  $\Xi$ , we have

$$\begin{aligned} \mathcal{R}_1(T) &\leq \sum_{t=1}^T \frac{1}{(\psi(S_t)^\top \theta^*)^2} \cdot \psi(S_t)^\top (\theta^* - \tilde{\theta}_t) \\ &\leq \sum_{t=1}^T \frac{1}{(\psi(S_t)^\top \theta^*)^2} \cdot \|\psi(S_t)\|_{H_t^{-1}(\theta^*; \lambda_t)} \cdot \|\theta^* - \tilde{\theta}_t\|_{H_t^{-1}(\theta^*; \lambda_t)}, \end{aligned}$$

where the first inequality is due to Eq. (31) and the second holds by Cauchy-Schwartz inequality. Notice that  $\tilde{\theta}_t, \theta^* \in C_t(\hat{\theta}_t; \delta, \lambda_t)$  under  $\Xi$ , we have

$$\|\theta^* - \tilde{\theta}_t\|_{H_t^{-1}(\theta^*; \lambda_t)} \leq 8\gamma_t(\delta, \lambda_t)$$

by Liu et al. (2024a, Lemma 30). Denote  $\gamma := \sup_{t \in [T]} \gamma_t(\delta, \lambda_t)$ , we can upper bound  $\mathcal{R}_1(T)$  by

$$\mathcal{R}_1(T) \leq 8 \cdot \sum_{t=1}^T \frac{1}{(\psi(S_t)^\top \theta^*)^2} \cdot \|\psi(S_t)\|_{H_t^{-1}(\theta^*; \lambda_t)} \cdot \gamma.$$

Denote  $A_t := \psi(S_t)/\psi(S_t)^\top \theta^*$ , we have  $H_t(\theta^*; \lambda) = \sum_{i < t} A_t^\top A_t + \lambda_t I$  and  $\|A_t\|_2 \leq \sum_{i \in S_t} \|\phi(i)\|_2$ .  $\ell^*(S_t) \leq KL^*$ . Then we have

$$\begin{aligned} \mathcal{R}_1(T) &\leq 8\gamma \sqrt{\sum_{t=1}^T \|A_t\|_{H_t^{-1}(\theta^*; \lambda_t)}^2} \cdot \sqrt{\sum_{t=1}^T \frac{1}{(\psi(S_t)^\top \theta^*)^2}} \\ &\leq 16\gamma \cdot \sqrt{d(1 + KL^*/\lambda) \cdot \log(1 + KL^*T/d\lambda)} \cdot \sqrt{\sum_{t=1}^T \frac{1}{(\psi(S_t)^\top \theta^*)^2}}, \end{aligned}$$

where the first inequality is due to the Cauchy-Schwartz inequality, and the second inequality is due to the elliptical potential lemma of Abbasi-Yadkori et al. (2011). Moreover, by Liu et al. (2024a, Lemma 31), we have

$$\begin{aligned} \sqrt{\sum_{t=1}^T \frac{1}{(\psi(S_t)^\top \theta^*)^2}} &\leq \sqrt{T \cdot \frac{1}{(\psi(S^*)^\top \theta^*)^2} + 2 \cdot \mathcal{R}(T)} \\ &\leq \sqrt{T \cdot (\ell^*(S^*))^2} + \sqrt{2 \cdot \mathcal{R}(T)}, \end{aligned}$$

which shows that for  $\lambda := \inf_t \lambda_t$ ,

$$\begin{aligned} \mathcal{R}_1(T) &\leq 16\gamma \cdot \sqrt{d(1 + L^*/\lambda) \cdot \log(1 + L^*T/d\lambda)} \cdot \sqrt{T \cdot (\ell^*(S^*))^2} \\ &\quad + 16\gamma \cdot \sqrt{d(1 + L^*/\lambda) \cdot \log(1 + L^*T/d\lambda)} \cdot \sqrt{2 \cdot \mathcal{R}(T)}. \end{aligned}$$

Next we give the upper bound for  $\mathcal{R}_2(T)$ . Recall that

$$\mathcal{R}_2(T) = \sum_{t=1}^T \frac{2}{\xi^3} \cdot (\psi(S_t)^\top \theta^* - \psi(S^*)^\top \theta^*)^2.$$

Then, under  $\Xi$ , we have

$$\begin{aligned}\mathcal{R}_2(T) &\leq \sum_{t=1}^T \frac{2}{\xi^3} \cdot \left\langle \psi(S_t), \theta^* - \tilde{\theta}_t \right\rangle^2 \\ &\leq \frac{2}{\ell^*(S^*)^3} \cdot \sum_{t=1}^T \|\psi(S_t)\|_{H_t^{-1}(\theta^*; \lambda_t)}^2 \cdot \|\theta^* - \tilde{\theta}_t\|_{H_t^{-1}(\theta^*; \lambda_t)}^2 \\ &\leq \frac{2}{\ell^*(S^*)^3} \cdot 64\gamma^2 \cdot \sum_{t=1}^T \|\psi(S_t)\|_{H_t^{-1}(\theta^*; \lambda_t)}^2,\end{aligned}$$

where the first inequality is according to Eq. (31), the second inequality is due to the Cauchy-Schwartz inequality, and the last inequality holds by Lemma B.1. Denote

$$\Lambda_t := \lambda_t I + \sum_{i < t} \psi(S_i)^\top \psi(S_i).$$

Then we have

$$\sup_{S \in \mathcal{S}} (\psi(S)^\top \theta^*) \cdot \Lambda_t^{-1} \succ H_t^{-1}(\theta^*; \lambda_t),$$

which further implies

$$\begin{aligned}\mathcal{R}_2(T) &\leq \frac{2}{\ell^*(S^*)^3} \cdot 64\gamma^2 \cdot \sup_{S \in \mathcal{S}} (\psi(S)^\top \theta^*) \cdot \sum_{t=1}^T \|\psi(S_t)\|_{\Lambda_t^{-1}}^2 \\ &\leq \frac{2}{\ell^*(S^*)^3} \cdot 64\gamma^2 \cdot \sup_{S \in \mathcal{S}} (\psi(S)^\top \theta^*) \cdot 2dL^* \log(1 + L^*T/d\lambda) \\ &= \frac{256}{\ell^*(S^*)^3} \sup_{S \in \mathcal{S}} (\psi(S)^\top \theta^*) \cdot \gamma^2 \cdot dL^* \log(1 + L^*T/d\lambda).\end{aligned}$$

Therefore, we have

$$\begin{aligned}\mathcal{R}(T) &\leq 16\gamma \cdot \sqrt{d(1 + L^*/\lambda) \cdot \log(1 + L^*T/d\lambda)} \cdot \sqrt{T \cdot (\ell^*(S^*))^2} \\ &\quad + 16\gamma \cdot \sqrt{d(1 + L^*/\lambda) \cdot \log(1 + L^*T/d\lambda)} \cdot \sqrt{2 \cdot \mathcal{R}(T)} \\ &\quad + \frac{256}{\ell^*(S^*)^3} \sup_{S \in \mathcal{S}} (\psi(S)^\top \theta^*) \cdot \gamma^2 \cdot dL^* \log(1 + L^*T/d\lambda) + 1.\end{aligned}$$

Notice that for  $x \leq A\sqrt{x} + B$ , we have  $x \leq 2A^2 + B$ . Therefore, we have

$$\begin{aligned}\mathcal{R}(T) &\leq 16\gamma \cdot \sqrt{dT} \cdot \sqrt{(\ell^*(S^*))^2(1 + L^*/\lambda) \cdot \log(1 + L^*T/d\lambda)} \\ &\quad + 256\gamma^2 \cdot dL^* \cdot \log(1 + L^*T/d\lambda) \cdot \left( \frac{\sup_{S \in \mathcal{S}} (\psi(S)^\top \theta^*)}{\ell^*(S^*)^3} + 2 \right) + 1, \\ &\leq \tilde{\mathcal{O}}(\sqrt{d^3 T})\end{aligned}$$

□