On the Sublinear Regret of Continuous K-Max Bandits

Yu Chen[*1] **Siwei Wang**[*2] **Longbo Huang**[1] **Wei Chen**[2]

[1]Institute for Interdisciplinary Information Sciences, Tsinghua University
[2]Microsoft Research Asia

Abstract

The K-Max combinatorial multi-armed bandit problem arises in applications such as recommendation and distributed decision making, where the reward is determined by the maximum outcome among K selected arms. When outcomes are continuous and only the maximum value together with the winner's index is observed, this problem introduces unprecedented difficulties including discretization errors, non-deterministic tie-breaking, and severe estimation biases. To overcome these barriers, we introduce `DCK-UCB`, an efficient algorithm combining adaptive discretization with bias-corrected confidence bounds. We prove that `DCK-UCB` achieves a $\widetilde{\mathcal{O}}(T^{3/4})$ regret bound, the *first* sublinear guarantee in this setting. Numerical experiments show strong performance over baseline methods. Furthermore, for the specific case of exponential distributions under full-bandit feedback, we propose the `MLE-Exp` algorithm that attains a near-optimal $\widetilde{\mathcal{O}}(\sqrt{T})$ regret bound. This work establishes fundamental theoretical guarantees and provides a powerful algorithmic solution for continuous combinatorial bandits.

1 INTRODUCTION

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 returns a reward signal according to the outcomes of the selected arms. As a popular variant, *K-Max Bandits* [Goel et al., 2006] focuses on the maximum outcome within a selected subset of K arms, naturally capturing real-world scenarios where the decision quality depends only on the best realized outcome.

In this paper, we consider the *Continuous K-Max Bandits* problem where the outcome for each arm follows a continuous distribution, motivated by the prevalence of continuous real-valued signals (e.g., job completion time in distributed computing, latency in server scheduling, bidding price in auctions, and ratings of selected products in online advertising). Moreover, we focus on a limited feedback environment that provides only the maximum outcome (reward) and the index of the winning arm (the arm that achieved the maximum). This novel framework, termed *Continuous K-Max Bandits* with *value-index feedback*, is applicable to various real-world scenarios. For example, 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. The latency we need to optimize is the shortest response time among the selected servers. Likewise, the seller invites a fixed number of bidders to a first-price auction; each invited bidder independently draws and submits a bid from an unknown continuous valuation distribution; the highest bid wins and is paid, and the seller observes only the winning bid and the winner's identity, while all other bids remain unobserved.

Tackling this problem is highly challenging. Most existing studies for K-Max bandits focus on discrete outcome settings [Simchowitz et al., 2016, Wang et al., 2024] or require feedback on every selected arm (semi-bandit feedback) [Chen et al., 2016a, Wang and Chen, 2017]. For the continuous K-Max bandits, naively applying existing discrete approaches can lead to significant discretization errors and impractical operations, hindering learning efficiency. Furthermore, the combination of continuous outcomes and limited *value-index feedback* introduces unique challenges from biased observations, which are not typically faced in

[*]Equal contribution. This work was done while Yu Chen was an intern at Microsoft Research Asia. Corresponding author: Wei Chen (weic@microsoft.com).

discrete settings or with richer feedback.

Specifically, existing studies face three critical limitations when tackling the continuous distribution and value-index feedback: First, most algorithms require semi-bandit feedback [Chen et al., 2016a, Simchowitz et al., 2016, Wang and Chen, 2017] revealing all selected arms' outcomes, while practical systems often only provide the winning arm's index and value, introducing bias since observations occur only when an arm wins. Second, greedy-style submodular bandit approaches [Streeter and Golovin, 2008, Fourati et al., 2024] typically compete with a $(1-1/e)$-approximation benchmark [Nemhauser et al., 1978], which can result in linear regret under our exact-optimum benchmark. Third, most prior solutions for K-Max bandits assume binary [Simchowitz et al., 2016] or finitely supported outcomes [Wang et al., 2024], struggling with continuous distributions under value-index feedback due to: (i) discretization error versus learning efficiency trade-offs, (ii) biased estimation from limited observability, and (iii) non-deterministic tie-breaking when several continuous values map to one discrete bin. We further analyze the difficulties and challenges in Section 4.

To fully address these issues, we design a novel framework for Continuous K-Max Bandits with value-index feedback through two key technical innovations: **(i)** We develop a principled approach to discretize the continuous outcome distributions, transforming the problem into a discrete K-Max bandit instance (Section 5.1). This allows us to control discretization error and utilize an efficient offline $(1-\epsilon)$-approximated optimization oracle (for any $\epsilon \in (0,1)$), crucially avoiding the linear regret associated with traditional greedy algorithms limited by $(1-1/e)$ approximations. **(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 (Section 5.2). In this paper, we design a bias-corrected estimation method coupled with a novel concentration analysis incorporating bias-aware error control (Section 5.3), enabling us to jointly manage estimation variance and discretization-induced bias to achieve sublinear regret.

Contributions. Our contribution can be summarized as follows:

(i) Our primary contribution is algorithm `DCK-UCB` (Algorithm 1 in Section 5), designed for K-Max bandits with general continuous distributions under value-index feedback. We prove that `DCK-UCB` achieves a regret upper bound of $\widetilde{\mathcal{O}}(T^{3/4})$ (Theorem 6), which is the first polynomial-time algorithm to achieve a sublinear regret guarantee for continuous K-Max bandits with value-index feedback. We also provide numerical experiments in Appendix B to empirically validate the effectiveness of `DCK-UCB`.

(ii) The most interesting technical innovation is the newly developed bias-corrected estimation in `DCK-UCB`. Specifically, this technique helps us conduct efficient learning under biased observations due to the nondeterministic tie-breaking arising from the discretization. This novel technique can be of independent interest.

(iii) We further develop the Maximum Likelihood Estimation (MLE) under a special case where outcomes for each arm follow exponential distributions. Our algorithm, `MLE-Exp` (Algorithm 2), designed for K-Min bandits (a variant of continuous K-Max), leverages the unique structure of exponential distributions to bypass the need for discretization and its associated biases. Combined with the lower bound $\Omega(\sqrt{T})$ proven in Appendix E, `MLE-Exp` achieves a near-optimal $\widetilde{\mathcal{O}}(\sqrt{T})$ regret (Theorem 7), representing a significant improvement over the general continuous case.

Notations. For any integer $n \geq 1$, we use $[n]$ to denote the set $\{1, 2, \ldots, n\}$. $\mathcal{O}$ is used to suppress all constant factors, while $\widetilde{\mathcal{O}}$ further suppresses all logarithmic factors. Bold letters such as $\boldsymbol{x}$ 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 σ-algebra generated by all random observations made within the first t time slots.

2 RELATED WORKS

The K-Max bandit problem departs from standard Combinatorial Multi-Armed Bandits (CMABs) [Cesa-Bianchi and Lugosi, 2012, Chen et al., 2013, 2014, Kveton et al., 2015b, Combes et al., 2015, Liu et al., 2023b], which suffices to learn the expected outcome of each arm. In contrast, K-Max bandits, whose reward is the maximum outcome among the selected arms [Goel et al., 2006, Gopalan et al., 2014], require knowing full distributions. In the following, we organize related work by feedback type.

Full-Bandit (Value) Feedback. Here the learner observes only the numerical reward, i.e., the maximum outcome over the chosen subset. Gopalan et al. [2014] derived a Thompson Sampling regret bound of $\mathcal{O}\left(\sqrt{\binom{N}{K}T}\right)$, but their analysis assumes a known finite parameter set and the regret scales exponentially in K. For pure exploration, Simchowitz et al. [2016] study the Bernoulli case. A submodular-maximization view yields $(1-1/e)$ approximation guarantees via greedy selection [Streeter and Golovin, 2008, Yue and Guestrin, 2011, Nie et al., 2022, Fourati et al., 2024]; K-Max naturally satisfies submodularity. However, such approximation-style guarantees lead to linear regret when the baseline is the *true* optimal subset. Achieving sublinear regret for general K-Max with full-bandit feedback therefore remains open.

Value–Index Feedback. The learner observes both the maximum outcome and the index of the winning arm. This resembles CMAB with probabilistic triggering [Wang and Chen, 2017, Liu et al., 2023b, 2024b]. Under this feedback model, Wang et al. [2024] analyze discrete K-Max with finite outcome support and deterministic tie-breaking and obtain $\widetilde{\mathcal{O}}(\sqrt{T})$ regret; Simchowitz et al. [2016] also study Bernoulli outcomes. However, these finite-support methods do not extend to the continuous setting considered in this paper due to non-zero discretization error and nondeterministic tie-breaking. DART [Agarwal et al., 2020] is another related value-index method that avoids finite-support enumeration via scalar arm-level estimates, but it relies on a structural ordering condition, informally that "good arms generate good actions": replacing a worse arm by a better one should improve the expected joint reward in every fixed context. Such a context-independent total order need not exist for general K-Max rewards, because $r^*(S) = \mathbb{E}[\max_{i\in S} X_i]$ depends on the full outcome distributions and the relative value of two arms can flip with the other arms in the selected set.

Semi-Bandit Feedback. Semi-bandit reveals the outcomes of *all* selected arms, enabling unbiased per-arm estimation and simpler UCB/TS-style learning. This richer feedback supports $\mathcal{O}(\sqrt{T})$ regret for discrete or continuous K-Max settings [Simchowitz et al., 2016, Jun et al., 2016, Chen et al., 2016a,b, Slivkins et al., 2019]. Nonetheless, these algorithms rely on observations strictly richer than value-only or value–index feedback and are not directly applicable to our setting.

3 FORMULATION

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 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[1]. 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:

$$r^*(S) := \mathbb{E}[\max\{X_i : i \in S\}].$$

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

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

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

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

4 CHALLENGES AND RESOLUTION IN CONTINUOUS K-MAX BANDITS WITH VALUE–INDEX FEEDBACK

To avoid learning over exponentially many super-actions, the key idea of the CMAB framework is to reconstruct per-arm information from the feedback, reducing the statistical burden to per-arm quantities [Chen et al., 2013]. However, for continuous K-Max bandits formulated in Section 3, the combination of continuous outcome distributions, the `max` operator, and value-index observability creates obstacles that make previous reduction methods for CMABs fail.

We outline the shortcomings of three existing methods: *(i)* discretization to finite-support settings; *(ii)* parametric Maximum Likelihood Estimation (MLE); and *(iii)* submodularity-driven greedy, and demonstrate how our solution overcomes these limitations.

4.1 DISCRETIZATION TO FINITE-SUPPORT SETTINGS

For finite-support K-Max bandits with value-index feedback and deterministic tie-breaking (i.e., the winner is deterministic when several arms have the same outcome), expanding each arm into binary sub-arms admits efficient algorithms and $\widetilde{O}(\sqrt{T})$ regret bounds [Wang et al., 2024]. In the continuous case, however, discretization creates artificial ties at the top bins, while value-index feedback reveals only a random winner in the same bin; the winner distribution under tie-breaking can vary across continuous outcome distributions (see Section 5.3 for more details). Hence, naive winner-frequency estimators become *selection biased*, which fails the existing analysis based on unbiased estimation of frequencies.

[1] Indeed, our theoretical guarantees and algorithmic approach naturally extend to all subsets S with $|S| \le K$. When the super arm set is not the set of all subsets of cardinality at most K, our approach still works as long as there exists an efficient oracle to solve the offline optimization problem.

Our Solution: We introduce a novel bias-corrected estimation method and develop the concentration analysis incorporating bias-aware error control (Section 5.3), enabling us to jointly manage estimation variance and discretization-induced bias to achieve efficient learning with biased estimators.

4.2 PARAMETRIC MAXIMUM LIKELIHOOD ESTIMATION

In this way, we often assume $\{f_i(\cdot;\theta_i), F_i(\cdot;\theta_i)\}_{i=1}^N$ to be the probability density function (PDF) and cumulative distribution function (CDF) for each arm, where the vector θ_i denotes the parameter for arm i. For round t with action S_t, the feedback (r_t, i_t) has joint density

$$p_\theta(r_t{=}r, i_t{=}i \mid S_t) = f_i(r;\theta_i) \prod_{j\in S_t\setminus\{i\}} F_j(r;\theta_j), \quad (2)$$

and the log-likelihood for round t is

$$L_t(\theta) = \log f_{i_t}(r_t;\theta_{i_t}) + \sum_{j\in S_t\setminus\{i_t\}} \log F_j(r_t;\theta_j). \quad (3)$$

To perform online learning analysis for MLE, it must be assumed that the $\{L_t(\theta)\}_t$ is at least embedded into a finite-dimensional statistical manifold. However, unlike Generalized Linear Bandits (GLB) [Liu et al., 2024a, Lee et al., 2024a] and Multinomial Logistic (MNL) Bandits [Liu et al., 2024b, Lee et al., 2024b], the log-likelihood function L_t for K-Max bandits *couples both PDFs and CDFs*, which derives inherent hardness for bounding the statistical dimension of $\{L_t(\theta)\}_t$ for general continuous distribution D_i.

Our Solution: We investigate a structured exponential family where each arm $i \in [N]$ follows an independent exponential distribution. In this case we design a likelihood-based estimator and achieve $\tilde{O}(\sqrt{T})$ regret (see Section 6).

4.3 SUBMODULARITY-DRIVEN GREEDY

Define $f(S) = \mathbb{E}[\max_{i\in S} X_i]$. For any realization $x \in \mathbb{R}^n$, the set function $S \mapsto \max_{i\in S} x_i$ is *monotone submodular*. Since nonnegative linear combinations preserve submodularity, f is monotone submodular.[2] For general monotone submodular objectives, greedy attains the tight $(1-1/e)$ offline ratio under a cardinality constraint [Nemhauser et al., 1978]. This approximation benchmark is therefore natural in submodular bandit work [Yue and Guestrin, 2011, Nie et al., 2022, Fourati et al., 2024]. However, we consider the regret Eq. (1) measured against the *true* optimum S^*. Then any α-approximate oracle with $\alpha < 1$ induces a per-round gap of at least $(1-\alpha)\, f(S^*)$, resulting in the *linear* regret.

[2]This observation also justifies the frequent use of greedy oracles for K-cardinality constraints.

Implication for our problem: The preceding argument demonstrates that a submodularity-driven greedy algorithm is not appropriate for minimizing regret relative to the true optimum, S^*, in our scenario. Furthermore, our numerical experiments reveal a linear increase in regret when using the Greedy algorithm (Appendix B).

Preview of our approach. Guided by the above arguments, we (i) discretize the outcome space and give a polynomial-time offline oracle for the discrete reduction (Section 5.1); (ii) construct bias-corrected estimators tailored to value-index feedback (Section 5.3) and provide the polynomial-time algorithm with the first sublinear regret bound (Algorithm 1 in Section 5.2); and (iii) prove the first $\widetilde{O}(T^{3/4})$ regret for general continuous K-Max Bandits (Section 5.4). Moreover, for exponential distribution families, we recover the nearly-optimal $\tilde{O}(\sqrt{T})$ (Section 6).

5 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 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]. Assumption 1 imposes a simple regularity condition on the continuous outcomes: probability mass over any interval is comparable to the interval length. Similar smoothness conditions are common in continuous bandits [Li et al., 2017, Wang et al., 2019, Liu et al., 2023a], and many compactly supported models, such as mixed uniforms and truncated smooth distributions, satisfy it. We use this assumption to keep the discretized bins controlled and to bound the bias caused by artificial ties.

5.1 THE DISCRETIZATION OF CONTINUOUS K-MAX BANDITS

To derive our algorithm, we first introduce a discretization approach for continuous K-Max Bandits. Discretization is a standard tool in continuum and Lipschitz bandits [Kleinberg, 2004, Magureanu et al., 2014, Nika et al., 2020]. Under value-index feedback, however, binning continuous outcomes creates artificial ties, and the observed winners no longer give unbiased estimates of the discretized distribu-

tion. Our framework uses discretization together with a bias correction term, while also enabling compact storage and computation of continuous distributions, conversion to a set of binary arms, and the use of a polynomial-time offline optimization oracle.

Discretization. Since it is complex to estimate the general continuous distributions, a natural idea is to perform *discretization* with granularity ϵ. 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 granularity ϵ: $\bar{X}_i = v_j$ if $X_i \in M_j$, where $M = \lceil 1/\epsilon \rceil$ 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 $M_{\le j} = \cup_{j' \le j} M_{j'}$ and $M_{\ge j} = \cup_{j' \ge j} M_{j'}$. For simplicity, we denote $p^*_{i,j}$ as the probability that X_i falls in M_j. That is, 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, the discretized problem $\bar{\mathcal{B}}$ only depends on the discrete probability set $\boldsymbol{p}^* = \{p^*_{i,j} : i \in [N], j \in [M]\}$. Moreover, we set $\bar{r}(S; \boldsymbol{p}^*)$ as the expected reward of an action S in discrete K-Max bandits under the probability set $\boldsymbol{p}^*$:

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

Since $\max_{i \in S}(\bar{X}_i) = v_j \Rightarrow \max_{i \in S}(X_i) \in M_j$, the key property $\mathbb{P}\left[\max_{i \in S}(\bar{X}_i) = v_j\right] = \mathbb{P}\left[\max_{i \in S}(X_i) \in M_j\right]$ gives an upper bound for the discretization error (see formal version in Lemma 8):

Lemma 2. $|r^*(S) - \bar{r}(S; \boldsymbol{p}^*)| \le \epsilon, \quad \forall S \in \mathcal{S}$.

Converting a discrete arm to a set of binary arms. Inspired by Wang et al. [2024], we decompose a discrete arm $\bar{X}_i$ into M independent random variables $\{\bar{Y}_{i,j}\}_{j \in [M]}$, which makes it much easier to understand the estimation method. Here $\bar{Y}_{i,j}$ takes value v_j with probability $q^*_{i,j}$ (which is determined by Eq. (4)) such that $\bar{X}_i = \max_{j \in [M]} \bar{Y}_{i,j}$:

$$q^*_{i,j} := \frac{p^*_{i,j}}{1 - \sum_{j' > j} p^*_{i,j'}},\ p^*_{i,j} = q^*_{i,j} \cdot \prod_{j' > j} (1 - q^*_{i,j'}). \quad (4)$$

This conversion allows us to estimate $\boldsymbol{q}^* = \{q^*_{i,j} : i \in [N], j \in [M]\}$ instead of $\boldsymbol{p}^*$. For any action $S \in \mathcal{S}$, define $\bar{r}_q(S; \boldsymbol{q})$ as the expected maximum reward of $\{\bar{Y}_{i,j}\}_{i \in S, j \in [M]}$ with probability set $\boldsymbol{q}$. Then we have

Lemma 3. *For any $\boldsymbol{p}$ and $\boldsymbol{q}$ satisfying Eq. (4), we have for any action $S \in \mathcal{S}$, $\bar{r}_q(S; \boldsymbol{q}) = \bar{r}(S; \boldsymbol{p})$.*

The formal version of this lemma is given in Lemma 9. Moreover, the function $\bar{r}_q$ is monotone with respect to $\boldsymbol{q}$, i.e.,

Lemma 4 (Wang et al. [2024, Lemma 3.1])**.** *For any $\boldsymbol{q}'$ and $\boldsymbol{q}$ such that $q'_{i,j} \ge q_{i,j}$ holds for any $i \in [N], j \in [M]$, we have $\bar{r}_q(S; \boldsymbol{q}') \ge \bar{r}_q(S; \boldsymbol{q})$ for any $S \in \mathcal{S}$.*

Polynomial approximation offline oracle. For any discrete K-Max bandits with probability set $\boldsymbol{p}$, we can apply the *PTAS* algorithm [Chen et al., 2016a] as a polynomial time offline $(1-\epsilon)$-approximation optimization oracle for any given $\epsilon \in (0, 1)$. Moreover, for any probability set $\boldsymbol{q}$, we can convert it to $\boldsymbol{p}$ by Eq. (4), input this $\boldsymbol{p}$ to the PTAS oracle and get the approximation solution $\mathrm{PTAS}(\boldsymbol{p})$ satisfying

$$\begin{aligned} &\bar{r}_q\left(\mathrm{PTAS}(\boldsymbol{p}); \boldsymbol{q}\right) = \bar{r}\left(\mathrm{PTAS}(\boldsymbol{p}); \boldsymbol{p}\right) \\ &\ge (1-\epsilon) \cdot \max_{S \in \mathcal{S}} \bar{r}(S; \boldsymbol{p}) = (1-\epsilon) \cdot \max_{S \in \mathcal{S}} \bar{r}_q(S; \boldsymbol{q}), \end{aligned} \quad (5)$$

which provides a way to control the relative error of the PTAS oracle.

5.2 DCK−UCB: THE EFFICIENT CONTINUOUS K-MAX BANDITS ALGORITHM

Building on the methodology in the previous subsection, we adapt the framework in Wang et al. [2024], and present DCK−UCB (Discretized Continuous K-Max with Upper Confidence Bounds, with pseudo-code in Algorithm 1), the first efficient algorithm addressing K-Max bandits with general continuous outcome distributions. At a high level, we first discretize the continuous K-Max bandits into discrete K-Max bandits. Then we convert every discrete arm to a set of binary arms and estimate the corresponding $\boldsymbol{q}^*$ parameters by $\bar{\boldsymbol{q}}$. Finally, we convert $\bar{\boldsymbol{q}}$ back to $\bar{\boldsymbol{p}}$, input $\bar{\boldsymbol{p}}$ to the PTAS oracle, and get the action we want to select.

In Line 3 of Algorithm 1, we calculate the optimistic estimator $\bar{q}^t_{i,j}$ which upper bounds $q^*_{i,j}$ with high probability. This is done by adding two upper confidence bonus terms $\beta^t_{i,j}$ and $(K-1)L^4/j^2$. Analysis shows that $\bar{q}^t_{i,j} \ge q^*_{i,j}$ with high probability (Lemma 5). The detailed discussion on this estimator will be given in the next subsection. In Lines 4-5, the agent converts this $\bar{\boldsymbol{q}}$ to $\bar{\boldsymbol{p}}$, and then runs the offline $(1-\epsilon)$-approximation optimization oracle PTAS 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 bin index j_t, i.e., $r_t \in M_{j_t}$. In Lines 7-8, the agent estimates $\boldsymbol{q}^*$ 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 $\tau \le t$ satisfying $i \in S_\tau$ and $j_\tau \le j$. Moreover, DCK−UCB achieves polynomial complexity, with $\mathrm{poly}(N, K, 1/\epsilon)$ per round time and $\mathcal{O}(N/\epsilon)$ space.

5.3 BIAS-CORRECTED ESTIMATORS

The key challenge in the algorithm design and theoretical analysis is that $\hat{q}^t_{i,j}$ is not an unbiased estimator for $q^*_{i,j}$. This means that except for the confidence radius due to the

Algorithm 1 DCK−UCB: Discretized Continuous K-Max Bandits with Upper Confidence Bonus

Require: Discretization granularity ϵ and the offline $(1-\epsilon)$-approximated optimization oracle PTAS for discrete K-Max bandits [Chen et al., 2016a].

1: Initialize $M \leftarrow \lceil 1/\epsilon \rceil$, $\hat{q}^1_{i,1} \leftarrow 1$ for every $i \in [N]$, and $\hat{q}^1_{i,j} \leftarrow 0$ for every $i \in [N], j > 1$.
2: **for** $t = 1, 2, \ldots, T$ **do**
3: For every $i \in [N], j \in [M]$, with confidence bonus $\beta^t_{i,j}$ defined in Eq. (7), we set

$$\bar{q}^t_{i,j} \leftarrow \min\left\{\hat{q}^t_{i,j} + \beta^t_{i,j} + (K-1)\frac{L^4}{j^2}, 1\right\}. \quad (6)$$

4: Convert $\bar{\boldsymbol{q}}^t$ to $\bar{\boldsymbol{p}}^t$ by Eq. (4).
5: Choose action $S_t \leftarrow \text{PTAS}(\bar{\boldsymbol{p}}^t)$.
6: Observe (r_t, i_t) by executing action S_t. Denote j_t as the bin index of r_t, i.e., $r_t \in M_{j_t}$.
7: For any $i, j \in [N] \times [M]$,

$$C_t(i,j) \leftarrow C_{t-1}(i,j) + \mathbb{1}\left[i = i_t \ \ \& \ \ j = j_t\right],$$
$$SC_t(i,j) \leftarrow SC_{t-1}(i,j) + \mathbb{1}\left[i \in S_t \ \ \& \ \ j \geq j_t\right].$$

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

randomness of the environment, we still need another bonus term to bound the bias to guarantee that $\bar{q}^t_{i,j}$ 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 $\hat{q}^t_{i,j} = C_t(i,j)/SC_t(i,j)$ is an unbiased estimator for $q^*_{i,j}$. This is because in this case,

$$\frac{C_t(i,j)}{SC_t(i,j)} = \frac{\#\text{ of } i_\tau = i \ \ \& \ \ j_\tau = j}{\#\text{ of } i \in S_\tau \ \ \& \ \ j_\tau \leq j} = \frac{\#\text{ of } X_i(\tau) \in M_j \ \ \& \ \ i \in S_\tau \ \ \& \ \ j_\tau \leq j}{\#\text{ of } i \in S_\tau \ \ \& \ \ j_\tau \leq j},$$

which is the fraction of $X_i(\tau) \in M_j$ conditioned on $i \in S_\tau, j_\tau \leq j$, and 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. [2024] who requires deterministic tie-breaking rule). When multiple arms have $X_i(\tau) \in M_j$, the observed winning arm $i_\tau = \arg\max X_i(\tau)$ is not a fixed one, and even we do not know the distribution of the winner. Because of this, we cannot guarantee that conditional on $i \in S_\tau, j_\tau \leq j$, we increase the counter $C_t(i,j)$ for every time step τ such that $X_i(\tau) \in M_j$. Some steps where $X_i(\tau) \in M_j$ 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 $\{\hat{q}^t_{i,j}\}$. Therefore, to guarantee that $\{\bar{q}^t_{i,j}\}$ is an upper confidence bound of $\{q^*_{i,j}\}$, we need another bonus term.

Specifically, this required bonus term arises because, conditional on $i \in S_\tau, j_\tau \leq j$, there are some time steps where $X_i(\tau) \in M_j$ but arm i is not the winner and thus we miss these steps in counter $C_t(i,j)$. When this happens, 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

$$\begin{aligned}
&\sum_{i' \neq i, i' \in S_\tau} \mathbb{P}[X_i(\tau) \in M_j, X_{i'}(\tau) \in M_j \mid i \in S_\tau, j_\tau \leq j] \\
=&\sum_{i' \neq i, i' \in S_\tau} \frac{p^*_{i,j} p^*_{i',j}}{\sum_{j' \leq j} p^*_{i,j'} \sum_{j' \leq j} p^*_{i',j'}} \\
\leq&\,(K-1)\frac{(L\epsilon)^2}{(j\epsilon/L)^2} = (K-1)\cdot(L^4/j^2),
\end{aligned}$$

where the last inequality follows from $p^*_{i,j} \leq L\epsilon$ and $\sum_{j' \leq j} p^*_{i,j'} \geq j\epsilon/L$ under Assumption 1. Thus Assumption 1 converts the artificial-tie probability into the explicit bias-correction term for every $q^*_{i,j}$.

Lemma 5. *Under Assumption 1, let the confidence radius be defined as*

$$\beta^t_{i,j} := \sqrt{8\frac{\log(NMT)}{SC_{t-1}(i,j)}}. \quad (7)$$

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

$$\left|\hat{q}^t_{i,j} - q^*_{i,j}\right| \leq \beta^t_{i,j} + (K-1)\cdot(L^4/j^2),$$

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

5.4 REGRET ANALYSIS AND DISCUSSION

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

Theorem 6. *Under Assumption 1, given the exploration bonus term $\beta_{i,j}^t$ in Eq. (7), 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 $\mathcal{R}(T) \leq \widetilde{\mathcal{O}}(L^2N^{1/4}K^{5/4}T^{3/4})$.*

Theorem 6 establishes a regret bound sublinear-in-T for our DCK-UCB algorithm. To the best of our knowledge, this is the first sublinear regret guarantee for the continuous K-Max bandit problem under the challenging value-index feedback model.

The key component of achieving Theorem 6 is the novel bias-corrected estimators and the concentration analysis with bias-aware error control, as detailed in Section 5.3. The analysis divides the cumulative regret into two parts: the estimation error which is upper bounded by the summation of the exploration bonus term $\beta_{i,j}^t$, and the bias-correction error which is upper bounded by the summation of the bias-correction term $(K-1)\cdot(L^4/j^2)$. The first one is upper bounded by $\widetilde{\mathcal{O}}\left(\sqrt{NKT}/\epsilon\right)$, following a regular CMAB-type analysis [Wang and Chen, 2017] where there are $NM \approx N/\epsilon$ bins in total and the player needs to pull $KM \approx K/\epsilon$ of them in each time step. The second term, on the other hand, needs to take summation over all the pulled $KM \approx K/\epsilon$ bins in every time step, while the bias-correction term for the j-th bin may influence the regret by at most $j\epsilon$. Thus, it is upper bounded by $\widetilde{\mathcal{O}}\left(K^2L^4T\epsilon\right)$. By choosing the discretization granularity ϵ for balancing these two terms, we achieve the regret guarantee in Theorem 6, which is the first theoretical result sublinear in T for continuous K-Max bandits. Please refer to Theorem 20 for the formal statement and complete proof.

Comparison to Prior Works. Our $\widetilde{\mathcal{O}}(T^{3/4})$ regret bound advances the state-of-the-art in several ways. While Wang et al. [2024] achieves $\mathcal{O}(S\sqrt{T})$ regret for discrete K-Max bandits with finite support size S, their approach does not work efficiently in the continuous case due to the infinite support size. Although using our discretization method can address the issue of infinite support size, their analysis still fails due to the non-deterministic tie-breaking problem and the biased estimators, which is one of the key contributions of our work.

Recent submodular bandits work [Pasteris et al., 2023, Fourati et al., 2024, Tajdini et al., 2024] reaches $\widetilde{\mathcal{O}}(T^{2/3})$ regret but has key limitations: their regret baseline is an $(1-1/e)$-approximation, which leads to a linear regret in our definition. Moreover, they require flexibility in subset size selection. Our framework overcomes these challenges through bias-corrected estimators with PTAS integration, effectively handling continuous K-Max bandits.

Discussion on the regret order in Theorem 6. The regret bounds established in Theorem 6 result from strategically selecting the discretization granularity ϵ to achieve an optimal balance between the estimation error and bias-correction error. The $\sqrt{NKT}$ factor appearing in the upper bound of exploration bonus term $\widetilde{\mathcal{O}}(\sqrt{NKT/\epsilon^2})$ aligns with established $\Omega(\sqrt{NKT})$ lower bounds in combinatorial multi-armed bandits Kveton et al. [2015b]. Regarding the upper bound of bias-correction error, our analysis in Section 5.3 shows that the estimation bias of $q_{i,j}^*$ is approximately the probability of another arm's outcome falling into the same discretized bin j, which leads to a bound of $\mathcal{O}(KL^4/j^2)$. Moreover, a positive bias in $q_{i,j}^*$ may cause the reward to be overestimated by at least ϵ. When aggregated across all K arms in the selected action S_t and all $t \leq T$, this yields the $\widetilde{\mathcal{O}}(K^2L^4T\epsilon)$ term, which seems inevitable.

In summary, we argue that these dependencies on N, K and T in both error components accurately capture the problem's fundamental complexity. Nevertheless, improving the relationship between discretization granularity ϵ and estimation error, e.g., reducing its dependency on ϵ from the current $\widetilde{\mathcal{O}}(\sqrt{NKT/\epsilon^2})$ to $\widetilde{\mathcal{O}}(\sqrt{NKT/\epsilon})$, could potentially enable a more favorable balance between these terms, leading to tighter bounds on N, K, and T. This is a promising yet challenging direction towards tighter regret bounds. Previous studies in combinatorial MABs [Liu et al., 2023b, 2024b] suggest a variance-adaptive analysis for the exploration bonus terms. However, achieving this faces substantial technical challenges, primarily because the observations are not i.i.d. due to the non-deterministic tie-breaking issue, which introduces new difficulties in deriving concentration inequalities for the variance terms of these biased observations.

6 BETTER PERFORMANCE UNDER THE EXPONENTIAL DISTRIBUTIONS

In this section, we demonstrate how specific distributional structure enables the improvement of the regret guarantee from $\widetilde{\mathcal{O}}(T^{\frac{3}{4}})$ to $\widetilde{\mathcal{O}}(\sqrt{T})$. We investigate the special case of K-Min exponential bandits, which can be viewed as a variant of continuous K-Max bandits where we seek to minimize losses rather than maximize rewards. For space limitations, we put the detailed discussion in Appendix C.

In the K-Min exponential bandits problem, each arm i generates a loss X_i from an exponential distribution $X_i \sim D_i = \mathrm{Exp}(\mu_i)$ where $\mu_i > 0$. The agent selects a subset $S_t \in \mathcal{S} = \{S \subseteq [N] : |S| = K\}$ and observes only the minimum loss $\ell_t = \min_{i\in S_t} X_i(t)$ (full bandit feedback without the winner's index).[3] Similar to previous works on general linear bandits [Liu et al., 2024a, Lee et al., 2024a, Liu et al., 2024b], we assume a linear structure where $\mu_i = \langle \phi(i), \theta^* \rangle$ for some unknown parameter $\theta^* \in \Theta \subset \mathbb{R}^d$

[3]This can be transformed into a K-Max problem by setting $Z_i(t) = -X_i(t)$ and viewing $Z_i(t)$ as rewards.

and a known bounded feature mapping $\phi : [N] \to \mathbb{R}^d$.

The critical insight that enables improved regret bounds is that the minimum of exponential random variables follows an exponential distribution with a parameter equal to the sum of the individual parameters, i.e.,

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

This property allows us to directly estimate θ^* using maximum likelihood estimation without requiring discretization or value-index feedback. Specifically, let $\psi(S) := \sum_{i \in S} \phi(i), \forall S \in \mathcal{S}$. Then with chosen action S_t and parameter θ, the observed loss should follow the exponential distribution $\text{Exp}\left(\sum_{i \in S} \phi(i)^T \theta\right) = \text{Exp}\left(\psi(S)^T \theta\right)$, whose probability density function is $f(x) = (\psi(S)^T \theta) \cdot e^{\left(-(\psi(S)^T \theta) x\right)}$. Because of this, the log-likelihood function is

$$L_t(\ell_t; S_t, \theta) : = \log\left(\psi(S_t)^\top \theta e^{\left(-\psi(S_t)^\top \theta \ell_t\right)}\right). \tag{8}$$

Note that this is a K-Min extension of Eq. (3) when the winner's index is not given. Specifically, let $f_\mu(x), F_\mu(x)$ be the PDF and CDF for exponential distribution with rate μ, Eq. (3) gives that $L_t(\theta) = \log(\sum_{i \in S_t} f_{\phi(i)^\top \theta}(\ell_t) \prod_{j \in S_t \setminus \{i\}} (1 - F_{\phi(j)^\top \theta}(\ell_t))) = \log(\sum_{i \in S_t} \phi(i)^\top \theta e^{-\phi(i)^\top \theta \ell_t} \prod_{j \in S_t \setminus \{i\}} e^{-\phi(j)^\top \theta \ell_t})$, which is exactly Eq. (8). This allows us to apply the MLE method. Denote $\mathcal{L}_t(\theta)$ as the summation of L_t and a regularization term with factor λ:

$$\mathcal{L}_t(\theta; \lambda) := \sum_{i<t} -L_i(\ell_i; S_i, \theta) + \frac{\lambda}{2}\|\theta\|^2, \tag{9}$$

Then we are ready to present the algorithm `MLE-Exp` (Algorithm 2), which constructs confidence sets around the MLE estimate and selects actions that minimize the expected loss.

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 λ_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. [2024a], Liu et al. [2024b], in Line 3, we construct a confidence set $C_t(\hat{\theta}_t; \delta)$ (defined in Eq. (10)), 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 $H_t(\theta) := \nabla^2_\theta \mathcal{L}_t(\theta)$:

$$C_t(\hat{\theta}_t; \delta) := \left\{\theta \in \Theta : \left\|g_t(\theta) - g_t(\hat{\theta}_t)\right\|_{H_t^{-1}(\theta)} \leq \gamma_t(\delta)\right\}, \tag{10}$$

Lemma 22 shows that with probability at least $1-\delta$, we have $\theta^* \in C_t(\hat{\theta}_t; \delta)$. Then in Line 4, we apply a double oracle to look for the action S whose expected loss under a parameter θ 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.

Algorithm 2 `MLE-Exp`: MLE for K-Min Exponential Bandits

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

1: **for** $t = 1, \ldots, T$ **do**

2: Compute MLE $\hat{\theta}_t$ by

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

where $\mathcal{L}_t(\theta; \lambda)$ is given in Eq. (9).

3: Construct the confidence set $C_t(\hat{\theta}_t; \delta, \lambda_t)$ according to Eq. (10)

4: Select action with minimum expected loss:

$$(S_t, \widetilde{\theta}_t) \leftarrow \operatorname*{argmax}_{S \in \mathcal{S}, \theta \in C_t(\hat{\theta}_t; \delta, \lambda_t)} \langle \psi(S), \theta \rangle.$$

5: Play action S_t and observe the loss ℓ_t.

6: **end for**

The following theorem shows that the `MLE-Exp` algorithm achieves a $\widetilde{O}(\sqrt{T})$ regret upper bound:

Theorem 7. *`MLE-Exp` achieves regret upper bound* $\mathcal{R}(T) \leq \widetilde{\mathcal{O}}\left(\sqrt{d^3 T}\right)$.

This theorem represents a significant improvement over the $\widetilde{\mathcal{O}}(T^{3/4})$ bound for general continuous K-Max bandits. Combined with the $\Omega(\sqrt{T})$ lower bound proven in Appendix E, we show that `MLE-Exp` achieves the nearly minimax optimal regret with respect to T, even without winner's index feedback. Note that the regret bound in Theorem 7 is derived under linear parametric assumptions, which explains why the bound depends on the feature dimension d rather than explicitly on N or K. This dimension d inherently encapsulates information about the problem structure, including the number of arms and selection size, which aligns with established results in other combinatorial linear bandit settings [Liu et al., 2023b, 2024b]. The detailed proof of Theorem 7 is provided in Appendix D.

Exponential K-Min bandits form a special case of continuous K-Max bandits. In Section 5, the general solution for continuous K-Max bandits reaches a regret bound of $\widetilde{\mathcal{O}}(T^{3/4})$ by handling discretization and bias correction under value-index feedback. In contrast, Theorem 7 uses the exponential structure to obtain a near-optimal regret bound even under weaker full-bandit feedback. This suggests that additional distributional structure can lead to sharper algorithms and guarantees.

7 CONCLUSION AND FUTURE WORK

We studied Continuous K-Max Bandits with value–index feedback, a challenging setting due to continuous outcomes, discretization error, and biased observations. We proposed `DCK-UCB`, the first efficient algorithm with a sublinear $\widetilde{\mathcal{O}}(T^{3/4})$ regret (Theorem 6), and validated it with numerical experiments (Appendix B). For the special case of exponential distributions, we introduced `MLE-Exp`, achieving a near-optimal $\widetilde{\mathcal{O}}(\sqrt{T})$ regret (Theorem 7) that matches the minimax lower bound (proven in Appendix E).

Future Directions: A key direction is improving the $\widetilde{\mathcal{O}}(T^{3/4})$ bound for general continuous distributions. Variance-aware algorithms, which exploit second-order statistics as in CMABs [Liu et al., 2023b, 2024b], may reduce regret to $\widetilde{\mathcal{O}}(T^{2/3})$. 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. Further avenues include establishing lower bounds for K-Max bandits and relaxing the bi-Lipschitz assumption.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

Mridul Agarwal, Vaneet Aggarwal, Christopher J. Quinn, and Abhishek Umrawal. DART: aDaptive Accept RejecT for non-linear top-K subset identification, 2020. URL `https://arxiv.org/abs/2011.07687`. Extended version of AAAI 2021 paper.

Nicolo Cesa-Bianchi and Gábor Lugosi. Combinatorial bandits. *Journal of Computer and System Sciences*, 78 (5):1404–1422, 2012.

Shouyuan Chen, Tian Lin, Irwin King, Michael R Lyu, and Wei Chen. Combinatorial pure exploration of multi-armed bandits. *Advances in neural information processing systems*, 27, 2014.

Wei Chen, Yajun Wang, and Siyu Yang. Efficient influence maximization in social networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 199–208, 2009.

Wei Chen, Yajun Wang, and Yang Yuan. Combinatorial multi-armed bandit: General framework and applications. In *International conference on machine learning*, pages 151–159. PMLR, 2013.

Wei Chen, Wei Hu, Fu Li, Jian Li, Yu Liu, and Pinyan Lu. Combinatorial multi-armed bandit with general reward functions. *Advances in Neural Information Processing Systems*, 29, 2016a.

Wei Chen, Yajun Wang, Yang Yuan, and Qinshi Wang. Combinatorial multi-armed bandit and its extension to probabilistically triggered arms. *Journal of Machine Learning Research*, 17(50):1–33, 2016b.

Richard Combes, Mohammad Sadegh Talebi Mazraeh Shahi, Alexandre Proutiere, et al. Combinatorial bandits revisited. *Advances in neural information processing systems*, 28, 2015.

Fares Fourati, Christopher John Quinn, Mohamed-Slim Alouini, and Vaneet Aggarwal. Combinatorial stochastic-greedy bandit. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 12052–12060, 2024.

Yi Gai, Bhaskar Krishnamachari, and Rahul Jain. Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations. *IEEE/ACM Transactions on Networking*, 20(5):1466–1478, 2012.

Ashish Goel, Sudipto Guha, and Kamesh Munagala. 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, 2006.

Aditya Gopalan, Shie Mannor, and Yishay Mansour. Thompson sampling for complex online problems. In *International conference on machine learning*, pages 100–108. PMLR, 2014.

David Janz, Shuai Liu, Alex Ayoub, and Csaba Szepesvári. Exploration via linearly perturbed loss minimisation. In *International Conference on Artificial Intelligence and Statistics*, pages 721–729. PMLR, 2024.

Kwang-Sung Jun, Kevin Jamieson, Robert Nowak, and Xiaojin Zhu. Top arm identification in multi-armed bandits with batch arm pulls. In *Artificial Intelligence and Statistics*, pages 139–148. PMLR, 2016.

Robert D. Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. In *Advances in Neural Information Processing Systems*, volume 17, 2004.

Branislav Kveton, Zheng Wen, Azin Ashkan, and Csaba Szepesvari. Combinatorial cascading bandits. *Advances in Neural Information Processing Systems*, 28, 2015a.

Branislav Kveton, Zheng Wen, Azin Ashkan, and Csaba Szepesvari. Tight regret bounds for stochastic combinatorial semi-bandits. In *Artificial Intelligence and Statistics*, pages 535–543. PMLR, 2015b.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Junghyun Lee, Se-Young Yun, and Kwang-Sung Jun. A unified confidence sequence for generalized linear models, with applications to bandits. *arXiv preprint arXiv:2407.13977*, 2024a.

Junghyun Lee, Se-Young Yun, and Kwang-Sung Jun. Improved regret bounds of (multinomial) logistic bandits via regret-to-confidence-set conversion. In *International Conference on Artificial Intelligence and Statistics*, pages 4474–4482. PMLR, 2024b.

Lihong Li, Yu Lu, and Dengyong Zhou. Provably optimal algorithms for generalized linear contextual bandits. In *International Conference on Machine Learning*, pages 2071–2080. PMLR, 2017.

Qinghua Liu, Praneeth Netrapalli, Csaba Szepesvari, and Chi Jin. 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, 2023a.

Shuai Liu, Alex Ayoub, Flore Sentenac, Xiaoqi Tan, and Csaba Szepesvári. Almost free: Self-concordance in natural exponential families and an application to bandits. *arXiv preprint arXiv:2410.01112*, 2024a.

Xutong Liu, Jinhang Zuo, Siwei Wang, John CS Lui, Mohammad Hajiesmaili, Adam Wierman, and Wei Chen. Contextual combinatorial bandits with probabilistically triggered arms. In *International Conference on Machine Learning*, pages 22559–22593. PMLR, 2023b.

Xutong Liu, Xiangxiang Dai, Xuchuang Wang, Mohammad Hajiesmaili, and John Lui. Combinatorial logistic bandits. *arXiv preprint arXiv:2410.17075*, 2024b.

Stefan Magureanu, Richard Combes, and Alexandre Proutiere. Lipschitz bandits: Regret lower bounds and optimal algorithms. *arXiv preprint arXiv:1405.4758*, 2014.

George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. An analysis of approximations for maximizing submodular set functions—i. *Mathematical programming*, 14:265–294, 1978.

Guanyu Nie, Mridul Agarwal, Abhishek Kumar Umrawal, Vaneet Aggarwal, and Christopher John Quinn. An explore-then-commit algorithm for submodular maximization under full-bandit feedback. In *Uncertainty in Artificial Intelligence*, pages 1541–1551. PMLR, 2022.

Andi Nika, Sepehr Elahi, and Cem Tekin. Contextual combinatorial volatile multi-armed bandit with adaptive discretization. In *International Conference on Artificial Intelligence and Statistics*, pages 1486–1496. PMLR, 2020.

Stephen Pasteris, Alberto Rumi, Fabio Vitale, and Nicolò Cesa-Bianchi. Sum-max submodular bandits. *arXiv preprint arXiv:2311.05975*, 2023.

Max Simchowitz, Kevin Jamieson, and Benjamin Recht. Best-of-k-bandits. In *Conference on Learning Theory*, pages 1440–1489. PMLR, 2016.

Aleksandrs Slivkins et al. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12 (1-2):1–286, 2019.

Matthew Streeter and Daniel Golovin. An online algorithm for maximizing submodular functions. *Advances in Neural Information Processing Systems*, 21, 2008.

Artin Tajdini, Lalit Jain, and Kevin G Jamieson. Nearly minimax optimal submodular maximization with bandit feedback. *Advances in Neural Information Processing Systems*, 37:96254–96281, 2024.

E Upfal and M Mitzenmacher. Probability and computing, 2005.

Qinshi Wang and Wei Chen. Improving regret bounds for combinatorial semi-bandits with probabilistically triggered arms and its applications. *Advances in Neural Information Processing Systems*, 30, 2017.

Yiliu Wang, Wei Chen, and Milan Vojnovic. Combinatorial bandits for maximum value reward function under value-index feedback. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=eMHn77ZKOp`.

Yining Wang, Ruosong Wang, Simon S Du, and Akshay Krishnamurthy. Optimism in reinforcement learning with generalized linear function approximation. *arXiv preprint arXiv:1912.04136*, 2019.

Yisong Yue and Carlos Guestrin. Linear submodular bandits and their application to diversified retrieval. *Advances in Neural Information Processing Systems*, 24, 2011.

On the Sublinear Regret of Continuous K-Max Bandits (Supplementary Material)

Yu Chen[*1] **Siwei Wang**[*2] **Longbo Huang**[1] **Wei Chen**[2]

[1]Institute for Interdisciplinary Information Sciences, Tsinghua University
[2]Microsoft Research Asia

[*]Equal contribution. This work was done while Yu Chen was an intern at Microsoft Research Asia. Corresponding author: Wei Chen (weic@microsoft.com).

A OMITTED PROOFS IN SECTION 5

In this section, we present the omitted proofs in Section 5, which include the full proof of Theorem 6.

A.1 DISCRETIZATION ERROR

First we show that the discretization from the original continuous problem $\mathcal{B}^*$ to $\bar{\mathcal{B}}$ with discretization width ϵ incurs a controllable error in expected loss, which is shown in Lemma 2 and formalized by the following lemma.

Lemma 8 (Formal version of Lemma 2)**.** *For any $S \in \mathcal{S}$, we have*

$$\bar{r}(S; \boldsymbol{p}^*) \le r^*(S) \le \bar{r}(S; \boldsymbol{p}^*) + \epsilon. \tag{11}$$

Proof. Notice that we have

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

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

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

where the inequality is given by the definition of M_j. This proves the left-hand side of Eq. (11). For the other side, we can similarly establish

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

□

A.2 CONVERTING TO BINARY ARMS

As detailed in Section 5.1, we set

$$q^*_{i,j} := \frac{p^*_{i,j}}{1-\sum_{j'>j} p^*_{i,j'}},\ p^*_{i,j} = q^*_{i,j} \cdot \prod_{j'>j}(1-q^*_{i,j'}),$$

which implies

$$q^*_{i,j} = \frac{p^*_{i,j}}{1-\sum_{j'>j} p^*_{i,j'}} = \frac{p^*_{i,j}}{\sum_{j'=1}^{j} p^*_{i,j'}} = \frac{\mathbb{P}[X_i \in M_j]}{\mathbb{P}[X_i \in M_{\le j}]}.$$

For any given probability set $\boldsymbol{q} = \{q_{i,j} : i \in [N], j \in [M]\}$, we can apply Eq. (4) to get the corresponding $\boldsymbol{p}$ defined as

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

Assume $\{Y^{\boldsymbol{q}}_{i,j}\}_{i\in[N],j\in[M]}$ is the set of independent binary random variables such that $Y^{\boldsymbol{q}}_{i,j}$ takes value $v_j = (j-1)\epsilon$ with probability $q_{i,j}$ and takes value 0 otherwise. Also let $\{X^{\boldsymbol{p}}_i\}_{i\in[N]}$ be the set of independent discrete random variables such that $X^{\boldsymbol{p}}_i$ takes value v_j with probability $p_{i,j}$ for every $j \in [M]$. Therefore, by simple calculation, $\max_{j\in[M]}\{Y^{\boldsymbol{q}}_{i,j}\}$ has the same distribution as $X^{\boldsymbol{p}}_i$.

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

$$\bar{r}_q(S;\boldsymbol{q}) = \sum_{j\in[M]} v_j \cdot \left(Q_j(S;\boldsymbol{q}) - Q_{j-1}(S;\boldsymbol{q})\right), \tag{12}$$

where we denote for simplicity

$$Q_j(S;\boldsymbol{q}) := \prod_{k\in S, j'>j}(1-q_{k,j'}). \tag{13}$$

$Q_j(S;\boldsymbol{q})$ 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 9 (Formal version of Lemma 3)**.** *For any $\boldsymbol{p} = \{p_{i,j}\}_{i\in[N],j\in[M]}$ and $\boldsymbol{q} = \{q_{i,j}\}_{i\in[N],j\in[M]}$ satisfying*

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

we have for any $S \in \mathcal{S}$,

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

Proof. Notice that by definition, we have

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

and

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

Notice that $\max_{j\in[M]}\{Y^{\boldsymbol{q}}_{i,j}\}$ has the same distribution as $X^{\boldsymbol{p}}_i$, we have

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

□

A.3 BIASED CONCENTRATION

We aim to use $\hat{q}^t_{i,j}$ to estimate $q^*_{i,j}$. However, this is a biased estimation. In this section, we carefully control the gap between the biased estimator $\hat{q}^t_{i,j}$ 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 a given action S_t in round t, the environment samples 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 considers 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 version of $X_i(t)$. 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 10 (Formal version of Lemma 5)**.** *Under event $\mathcal{E}_0$, we have for every $t \in [T]$ and $(i,j) \in [N] \times [M]$,*

$$\left|\hat{q}^t_{i,j} - q^*_{i,j}\right| \le \sqrt{8\frac{\log(NMT)}{SC_t(i,j)}} + (K-1)\cdot(L^4/j^2),$$

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

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

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

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

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

which implies that

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

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

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

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

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

By the union bound, we have

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

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 $\mathbb{P}[\neg \mathcal{E}_1] \le T^{-2}$.

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

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

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

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

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

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

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

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

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

Therefore, we have

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

□

A.4 OPTIMISTIC ESTIMATION

Lemma 11. *For $\beta_{i,j}^t$ given in Eq. (7), 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

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

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 10, we have under $\mathcal{E}_0$ 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. (7) 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 $S_t \leftarrow \mathrm{PTAS}(\bar{\boldsymbol{p}}^t)$ where $\bar{\boldsymbol{p}}^t$ is converted from $\bar{\boldsymbol{q}}^t$ by Eq. (4). Then by Lemmas 3 and 4, we have

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

□

A.5 REGRET DECOMPOSITION

Lemma 12. *Denote $Q_j^*(S_t) := \prod_{k \in S_t, j' > j}(1 - q_{k,j'}^*)$. We have*

$$|\bar{r}_q(S_t; \bar{\boldsymbol{q}}^t) - \bar{r}_q(S_t; \boldsymbol{q}^*)| \leq 2 \sum_{i \in S_t, j \in [M]} Q_j^*(S_t) \cdot v_j \cdot \left|\bar{q}_{i,j}^t - q_{i,j}^*\right|. \tag{14}$$

Proof. This lemma is given by directly applying Lemma 3.3 in Wang et al. [2024] by the definition of $\bar{r}_q$ in Eq. (12). □

Lemma 13. *Under Assumption 1, we can bound the regret of Algorithm 1 by*

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

where $\texttt{Bonus}_t$ and $\texttt{Bias}_t$ are defined by

$$\texttt{Bonus}_t := 4 \sum_{i \in S_t, j \in [M]} Q_j^*(S_t) \cdot v_j \cdot \beta_{i,j}^t, \tag{15}$$

and

$$\texttt{Bias}_t := 4 \sum_{i \in S_t, j \in [M]} Q_j^*(S_t) \cdot v_j \cdot (K-1)\frac{L^4}{j^2}. \tag{16}$$

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

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

By Lemma 2, we have

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

Then we have

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

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 and 11, we have

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

Then with Lemma 9, we have

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

where the last inequality holds by $\epsilon\bar{r}_q(S^*; \boldsymbol{p}^*) \leq \epsilon$ and $\mathbb{P}[\neg\mathcal{E}_1] \leq T^{-2}$ shown in Lemma 10.

Therefore, applying Lemma 12, we get

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

where $\texttt{Bonus}_t$ and $\texttt{Bias}_t$ are defined in Eqs. (15) and (16). □

A.6 BOUNDING THE BONUS TERMS

We apply similar methods in Wang and Chen [2017], Liu et al. [2023b] to give the bounds of $\sum_t \texttt{Bonus}_t$. We first give the following definitions.

Definition 14 (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 15 (Wang and Chen [2017, Definition 6]). *For each group S_j^l (Definition 14), 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 16 (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 14) identified by binary arm (i,j) and $1 \le l \le l_{i,j}^{\max}$, as long as*

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

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

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

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

Lemma 17 (Wang and Chen [2017, Lemma 4]). *For a series of integers $\{l_{i,j}^{\max}\}_{i\in[N],j\in[M]}$,*

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

for every round $t \ge 1$.

Proof. We prove this lemma by showing $\mathbb{P}[N_{t-1}^l(i,j)=s, SC_{t-1}(i,j) \le \frac{1}{3}N_{t-1}^l(i,j)\cdot 2^{-l}] \le t^{-3}$, for any fixed s with $0 \le s \le t-1$ and $\sqrt{\frac{8\log(NMT)}{\frac{1}{3}s\cdot 2^{-l}}} \le 1$. Let t_k be the round that $N^l(i,j)$ is increased for the k-th time, for $1 \le k \le s$. Let $Z_k = \mathbb{1}[S_{t_k} \ni i, j_{t_k} \le j]$ 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 $S_{t_k} \in S_j^l$, $\mathbb{E}[Z_k \mid Z_1,\dots,Z_{k-1}] \ge 2^{-l}$. Let $Z = Z_1+\cdots+Z_s$. By multiplicative Chernoff bound [Upfal and Mitzenmacher, 2005], we have

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

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

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

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. □

Lemma 18. *For given constant C, we have*

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

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

$$l_{i,j}^{\max} := \left\lceil \log_2 \frac{16KM}{C} \right\rceil, \quad \forall (i,j) \in [N] \times [M], \tag{17}$$

and for every integer l,

$$\kappa_{l,T}(C,s) := \begin{cases} 2 \cdot 2^{-l} & s = 0 \\ \sqrt{96 \cdot 2^{-l} \log(NMT)/s} & 1 \le s \le B_{l,T}(C) \\ 0 & s > B_{l,T}(C) \end{cases}, \tag{18}$$

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

$$B_{l,T}(C) := \left\lfloor 6144 \cdot 2^{-l} K^2 M^2 \log(NMT)/C^2 \right\rfloor. \tag{19}$$

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

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

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

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

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

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

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

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

and

$$\begin{aligned} Q_j^*(S_t) \cdot \min\left\{\beta_{i,j}^t, 1\right\} &\le 2 \cdot 2^{-l_{i,j}} \cdot \min\left\{ \sqrt{\frac{8\log(NMT)}{\frac{1}{3}N_{t-1}^{l_{i,j}}(i,j) \cdot 2^{-l_{i,j}}}}, 1 \right\} \\ &\le \min\left\{ \sqrt{\frac{96 \cdot 2^{-l_{i,j}} \log(NMT)}{N_{t-1}^{l_{i,j}}(i,j)}}, 2 \cdot 2^{-l_{i,j}} \right\}. \end{aligned} \tag{20}$$

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

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

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

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

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

Otherwise, for $1 \le N_{t-1}^{l_{i,j}}(i,j) \le B_{l_{i,j},T}(C)$, we have $Q_j^*(S_t) \cdot \min\left\{\beta_{i,j}^t, 1\right\} \le \kappa_{l_{i,j},T}(C, N_{t-1}^{l_{i,j}}(i,j))$ by Eqs. (18) and (20). Therefore, we get

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

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

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

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

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

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

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

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

where the first inequality is due to $\texttt{Bonus}(t) \leq 1$ and definition, and the second one is due to Lemma 17. 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. (19), 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

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

□

A.7 BOUNDING THE BIAS TERMS

Lemma 19. *Under Assumption 1, we have*

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

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

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

Therefore, we have

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

□

A.8 PROOF OF THEOREM 6

Theorem 20 (Formal version of Theorem 6)**.** *By setting $\beta_{i,j}^t$ in Eq. (7) and $\epsilon < 1/2$, we can control the regret of Algorithm 1 under Assumption 1 by*

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

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

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

Proof. By Lemma 13, we have

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

Take constant C as

$$C := \sqrt{\frac{NKM^2\log(NMT)}{T}}. \tag{21}$$

Then Lemma 18 shows that

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

Lemma 19 demonstrates that

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

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

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

which finishes the proof. □

B NUMERICAL EXPERIMENTS FOR DCK−UCB

We conduct numerical experiments to validate the effectiveness of our proposed DCK−UCB algorithm.

B.1 COMPARABLE BASELINES

We compare DCK−UCB with two baseline algorithms: **(1) Naive UCB**: A naive adaptation of UCB Lattimore and Szepesvári [2020] that treats each subset of K base arms (i.e., each super arm) as an abstract arm, with $\binom{N}{K}$ super arms in total. **(2) Submodular Greedy**: A submodularity-driven greedy algorithm (widely used for submodular maximization task Nie et al. [2022], Tajdini et al. [2024]) that exploits the submodularity of K-Max bandits, i.e., it repeats the process that uniformly explores available base arms and greedily includes the best one as part of its solution, until the solution is of size K.

We highlight that standard baselines for K-Max bandits, such as SDCB [Chen et al., 2016a] and CUCB [Wang et al., 2024], are not suitable for continuous K-Max bandits with value-index feedback. Specifically, SDCB [Chen et al., 2016a] requires semi-bandit feedback, which demands observations from every selected arm. In our case, the agent receives only the maximum outcome with the winner's index. CUCB [Wang et al., 2024] works in finite-support distribution settings by recording and estimating each observed outcome value. For continuous K-Max bandits, this would require unbounded memory and is therefore not directly applicable.

B.2 INSTANCE SETUP

We first construct the instance with $(N = 10, K = 5)$ and $(N = 12, K = 3)$, where each arm's outcome is piecewise uniform in $[0, 1]$ with bi-Lipschitz constant $L = 3$.

Instance generation. All synthetic instances are built from independent *piecewise-uniform* base arms on $[0, 1]$. This family gives a controlled continuous testbed: the bounded densities satisfy the bi-Lipschitz condition in Assumption 1, while random breakpoints and segment heights create heterogeneous, nonparametric shapes beyond a single simple parametric family. Piecewise-constant densities also make the ground-truth rewards $\mathbb{E}[\max_{i\in S} X_i]$ and optimal subsets numerically stable to compute, reducing benchmark-construction noise in the reported regret. For given (N, A, K, L), we construct N base arms $X_1, \ldots, X_N$ as follows. For each arm i, we first draw the number of breakpoints B_i uniformly from $\{\lfloor K/2 \rfloor + 1, \ldots, K\}$ and sample B_i i.i.d. points in $(0, 1)$, which are sorted into $0 = b_{i,0} < b_{i,1} < \cdots < b_{i,B_i} < b_{i,B_i+1} = 1$. On each segment $[b_{i,j}, b_{i,j+1})$ we draw a density level $p_{i,j} \sim \mathrm{Unif}[1/L, L]$ and normalize $\{p_{i,j}\}$ so that the piecewise-constant pdf integrates to 1. Hence each arm has a piecewise-linear cdf F_i and a piecewise-constant pdf (implementation in our generator).

Avoid Trivial Cases. Our greedy baseline adds one arm at a time by maximizing the marginal gain in $\mathbb{E}[\max_{i\in S} X_i]$ (ties broken by smaller index). To avoid trivial cases, we *reject* any randomly drawn instance unless this greedy rule is suboptimal:

$$\mathbb{E}\left[\max_{i\in S^\star} X_i\right] - \mathbb{E}\left[\max_{i\in S_{\text{greedy}}} X_i\right] \geq \texttt{gap},$$

where gap is a user-specified threshold. Concretely, we repeatedly resample the N arms (up to a fixed maximum number of attempts) until the above inequality holds, at which point the instance is accepted. Figure 1 summarizes the two instances used in our experiments via per-arm density heatmaps.

Code and instance generator are provided in the supplementary material.

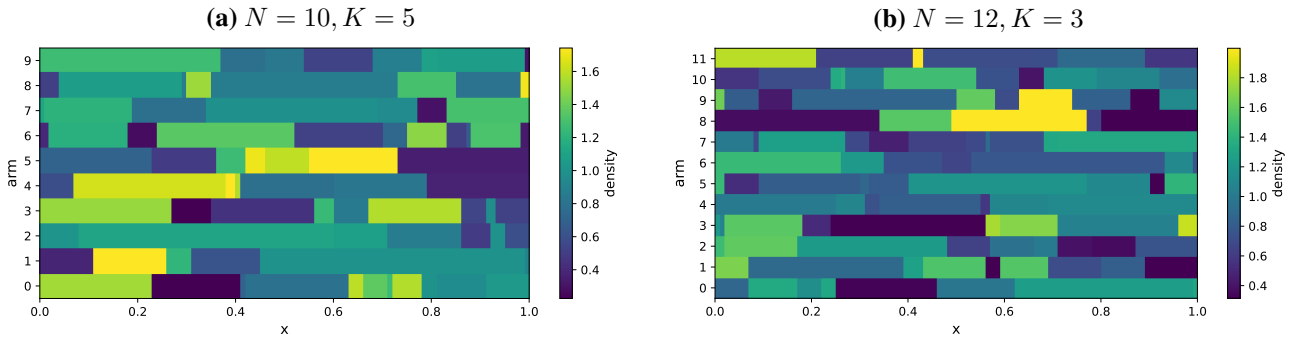

Figure 1: Per-arm density heatmaps for two benchmark instances. Rows correspond to arm indices and the horizontal axis is $x \in [0, 1]$; color encodes the probability density. Brighter cells indicate higher mass near that x.

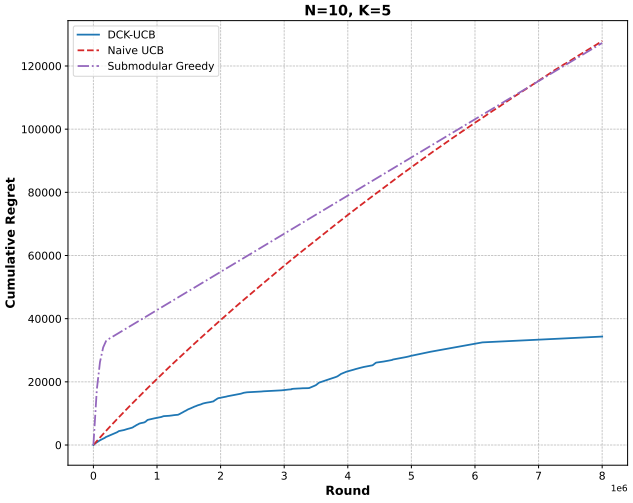

Figure 2: Comparison of cumulative regret for different algorithms on the problem instance with $N = 10$ and $K = 5$.

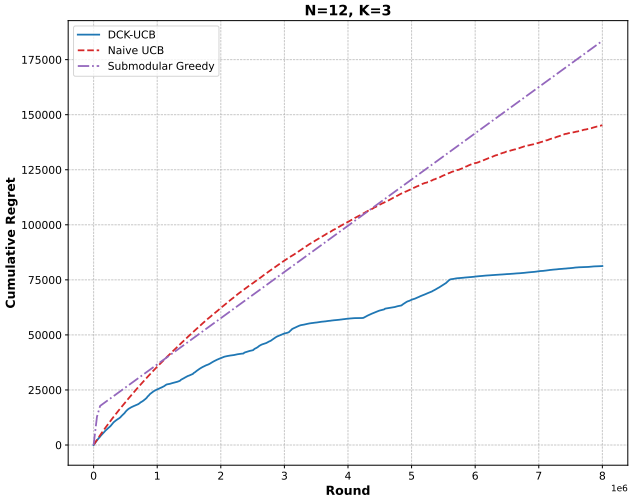

Figure 3: Comparison of cumulative regret for different algorithms on the problem instance with $N = 12$ and $K = 3$.

B.3 EXPERIMENT RESULTS

From Figures 2 and 3, we can observe that our proposed DCK−UCB algorithm consistently outperforms the baseline algorithms across different problem settings. In Figure 2, where we have $N = 10$ arms and need to select $K = 5$ arms in each round, DCK−UCB achieves significantly lower cumulative regret compared to both the Submodular Greedy and Naive UCB approaches.

The linear growth in regret observed for the Submodular Greedy algorithm stems from evaluating regret against the true optimal subset rather than an approximation benchmark. In our experimental setup, the optimal action consistently outperforms the greedy selection, creating a persistent gap that accumulates linearly over time.

The Naive UCB approach performs particularly poorly in this setting, likely due to the large number of super arms ($\binom{10}{5} = 252$ in total).

Similarly, in Figure 3, which presents a different configuration with $N = 12$ arms and $K = 3$, our algorithm maintains its strong performance. The Naive UCB approach performs better in this setting compared to the previous configuration, exhibiting a sublinear growth in regret with respect to round t. This improved performance can be attributed to the smaller number of super arms ($\binom{12}{3} = 220$ super arms compared to $\binom{10}{5} = 252$ in the previous configuration). Nevertheless, DCK−UCB still achieves lower cumulative regret.

These experimental results support our theoretical findings and demonstrate that DCK−UCB effectively balances the exploration-exploitation trade-off while handling the continuous nature of the arm distributions and the combinatorial structure of base arms.

C DETAILED EXPLANATION ON `MLE-EXP` FOR EXPONENTIAL K-MIN BANDITS

C.1 MOTIVATION

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 μ_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 μ_i, which allows incorporating features like distance, traffic, or weather conditions into the model.

C.2 THE K-MIN EXPONENTIAL BANDITS

Motivated by this, 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., $X_i \sim D_i = \mathrm{Exp}(\mu_i)$ where $\mu_i > 0$ is the parameter of arm i. Moreover, we assume that there exists a d-dimensional unknown parameter $\theta^* \in \mathbb{R}^d$ and a known feature mapping $\phi : [N] \to \mathbb{R}^d$ such that $\mu_i = \langle \phi(i), \theta^* \rangle$ holds for any $i \in [N]$. The feature mapping ϕ satisfies that $\|\phi(i)\|_2 \leq 1$ and the unknown parameter θ^* 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 $\ell^*(S) := \mathbb{E}[\ell_t \mid S]$ be the expected loss for action $S \in \mathcal{S}$, we further denote the best action $S^* = \mathrm{argmin}_{S \in \mathcal{S}} \ell^*(S)$ 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 $\ell_t = \min_{i \in S_t} X_i(t) = \min_{i \in S_t} -Z_i(t) = -\max_{i \in S_t} Z_i(t) = -r_t$. Thus, 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, which makes K-Min exponential bandits even more challenging.

C.3 ALGORITHM

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

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

Therefore, it becomes much easier to estimate the true parameter θ^* 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 θ, the observed loss should follow the exponential distribution $\mathrm{Exp}\left(\sum_{i \in S} \phi(i)^T \theta\right) = \mathrm{Exp}\left(\psi(S)^T \theta\right)$, whose probability density function is $f(x) = (\psi(S)^T \theta) \cdot e^{\left(-(\psi(S)^T \theta) x\right)}$. Because of this, the log-likelihood function is

$$L_t(\ell_t; S_t, \theta) : = \log\left(\psi(S_t)^\top \theta e^{\left(-\psi(S_t)^\top \theta \ell_t\right)}\right). \tag{22}$$

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

$$\mathcal{L}_t(\theta; \lambda) := \sum_{i<t} -L_i(\ell_i; S_i, \theta) + \frac{\lambda}{2}\|\theta\|^2, \tag{23}$$

Algorithm 3 `MLE-Exp`: MLE for K-Min Exponential Bandits

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

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

2: Compute MLE $\hat{\theta}_t$ by

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

where $\mathcal{L}_t(\theta; \lambda)$ is given in Eq. (23).

3: Construct the confidence set $C_t(\hat{\theta}_t; \delta, \lambda_t)$ according to Eq. (26)

4: $(S_t, \widetilde{\theta}_t) \leftarrow \operatorname{argmax}_{S\in\mathcal{S}, \theta\in C_t(\hat{\theta}_t;\delta,\lambda_t)} \langle \psi(S), \theta\rangle$

5: Play action S_t and observe the loss ℓ_t.

6: **end for**

where λ is the regularization factor. Then we present the algorithm `MLE-Exp` for K-Min exponential bandits in Algorithm 3.

In Line 2 of Algorithm 3, 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 λ_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. [2024a], Liu et al. [2024b], in Line 3, we construct a confidence set $C_t(\hat{\theta}_t; \delta)$ (defined in Eq. (26)), 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 $H_t(\theta) := \nabla^2_\theta \mathcal{L}_t(\theta)$. Then in Line 4, we apply a double oracle to look for the action S whose expected loss under a parameter θ 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.

Theorem 23 gives a formal version of Theorem 7, which achieves the first nearly-minimax optimal regret bound $\widetilde{\mathcal{O}}(\sqrt{T})$ for K-Min exponential bandits. This result advances the general result Theorem 6 in three aspects.

(1) We utilize the property of exponential distributions in K-Min bandits and achieve efficient unbiased estimation of the true parameter θ^* without the index feedback required in Algorithm 1 for general continuous distributions.

(2) We consider maximum log-likelihood estimator $\hat{\theta}_t$ and achieve the following concentration lemma inspired by the analysis of general linear bandits [Lee et al., 2024a, Liu et al., 2024a] and logistics bandits [Liu et al., 2024b].

Lemma 21. *With probability at least $1-\delta$, set λ_t in Eq. (24) and γ_t in Eq. (25), we have $\theta^* \in C_t(\hat{\theta}_t; \delta)$ holds for every $t \in [T]$.*

Equipped with the efficient MLE, we sharpen the cumulative regret bound from $\widetilde{\mathcal{O}}(T^{3/4})$ in Theorem 6 to $\widetilde{\mathcal{O}}(\sqrt{T})$.

(3) Thanks to the unbiased estimation of θ^*, we analyze the regret bound with the Hessian matrix $H_t(\theta)$ and avoid the inverse Lipschitz factor in Assumption 1, which might be infinite for exponential distributions.

D OMITTED PROOFS IN SECTION 6

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

We have

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

D.1 CONCENTRATION ARGUMENT FOR MLE

Lemma 22 (MLE Concentration). *For $L^* := \sup_{S\in\mathcal{S}} \ell^*(S)$, $M_1 := 2L^*$, and $V = \sup\{\|\theta\|_2 : \theta \in \Theta\}$, set*

$$\lambda_t := \max\left\{1, \frac{2dM_1}{V}\cdot\log\left(e\sqrt{1+\frac{tL^*}{d}}+\frac{1}{\delta}\right)\right\}, \tag{24}$$

and

$$\gamma_t(\delta,\lambda_t) := \sqrt{\lambda_t}\left(\frac{1}{2M_1}+V\right)+\frac{2M_1 d}{\sqrt{\lambda_t}}\left(\log(2)+\frac{1}{2}\log\left(1+\frac{tL^*}{\lambda_t d}\right)\right)+\frac{2M_1}{\sqrt{\lambda_t}}\log(1/\delta). \tag{25}$$

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

$$\theta^* \in C_t(\hat{\theta}_t;\delta,\lambda_t) := \left\{\theta\in\Theta : \left\|g_t(\theta;\lambda_t)-g_t(\hat{\theta}_t;\lambda_t)\right\|_{H_t^{-1}(\theta;\lambda_t)} \le \gamma_t(\delta,\lambda_t)\right\}, \tag{26}$$

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 Ξ.

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

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

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

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

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 $\varepsilon_t := \ell_t - \mathbb{E}[\ell_t\mid\mathcal{H}_t] = \ell_t - 1/(\psi(S_t)^\top\theta^*)$. We then 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 such that $|s|\le 1/M_1$. Since $\ell^*(S_t) = 1/(\psi(S_t)^\top\theta^*) \le L^*$, this range implies $s < \psi(S_t)^\top\theta^*$, so the moment below is well defined. We have

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

and by calculation,

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

where $a_t^* := \psi(S_t)^\top \theta^*$ and $x := s/a_t^*$. Since $1/a_t^* = \ell^*(S_t) \le L^*$, the choice $M_1 = 2L^*$ implies $|x| \le 1/2$ whenever $|s| \le 1/M_1$. For $|x| \le 1/2$, we have $-x - \log(1-x) \le x^2$, and hence

$$\mathbb{E}[\exp(s\varepsilon_t) \mid \mathcal{H}_t] \le \exp\left(\frac{s^2}{(a_t^*)^2}\right).$$

Denote $\nu_{t-1} := 1/(\psi(S_t)^\top \theta^*)^2$. Then, for $|s| \le 1/M_1$,

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

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

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

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

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

Therefore, for

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

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

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

holds for any $t \in [T]$. □

D.2 PROOF OF THEOREM 7

Theorem 23 (Formal version of Theorem 7)**.** *By setting $\delta = 1/T$, $\gamma_t(\delta)$ according to Eq. (25), and λ_t according to Eq. (24), Algorithm 2 enjoys the following regret guarantee:*

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

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

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

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

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 $\xi \in [\ell^*(S_t), \sup_t \ell^*(S_t)]$,

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

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

$$\psi(S^*)^\top \theta^* \le \psi(S_t)^\top \widetilde{\theta}_t. \tag{27}$$

Under Ξ, we have

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

where the first inequality is due to Eq. (27) and the second holds by Cauchy-Schwartz inequality. Notice that $\widetilde{\theta}_t, \theta^* \in C_t(\hat{\theta}_t; \delta, \lambda_t)$ under Ξ, we have

$$\left\|\theta^* - \widetilde{\theta}_t\right\|_{H_t^{-1}(\theta^*;\lambda_t)} \le 8\gamma_t(\delta, \lambda_t)$$

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

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

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

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

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

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

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

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

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

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

Then, under Ξ, we have

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

where the first inequality is according to Eq. (27), the second inequality is due to the Cauchy-Schwartz inequality, and the last inequality holds by Lemma 22. Denote

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

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

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

Therefore, we have

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

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

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

□

E REGRET LOWER BOUND FOR K-MIN EXPONENTIAL BANDITS

For completeness, we provide the regret lower bound for K-Min Exponential Bandits under full-bandit feedback. This lower bound, $\Omega(\sqrt{T})$, illustrates that `MLE-EXP` is nearly optimal, matching the regret upper bound $\widetilde{O}(\sqrt{T})$. The construction and proof of the lower bound are standard.

Theorem 24 (Regret lower bound for K-Min Exponential Bandits)**.** *Consider the K-Min Exponential Bandits defined in Section 6 with $N \geq K+1$ arms. There exists a constant $c > 0$ and a problem instance such that any learning algorithm satisfies*

$$\mathcal{R}(T) \;\geq\; c\sqrt{T}.$$

Proof. We construct the hard-to-learn instance with two close actions and apply the standard method in Lattimore and Szepesvári [2020] to establish the lower bound.

Fix $m > 0$ and a small parameter $\Delta \in (0, m)$. Construct $K-1$ base arms $B = \{1, \ldots, K-1\}$ each with rate $\mu_i = m, i \in B$, and two "contender" arms a and b whose rates differ across two environments:

$$\begin{aligned} \mathsf{P}: &\quad \mu_a = m + \Delta, \;\; \mu_b = m, \\ \mathsf{Q}: &\quad \mu_a = m, \;\; \mu_b = m + \Delta. \end{aligned}$$

Any valid action must select all arms in B and exactly one of $\{a, b\}$. Thus there are two actions: $S^a = B \cup \{a\}$ and $S^b = B \cup \{b\}$. Under P, S^a is optimal; under Q, S^b is optimal.

Notice that $\min_{i \in S} X_i \sim \mathrm{Exp}(\sum_{i \in S} \mu_i)$. Hence

$$\mu_{S^a} = \sum_{i \in S^a} \mu_i = Km + \Delta, \qquad \mu_{S^b} = \sum_{i \in S^b} \mu_i = Km.$$

The one-step regret gap is

$$\Delta_{\mathrm{gap}} = \ell^\star(S^b) - \ell^\star(S^a) = \frac{1}{Km} - \frac{1}{Km + \Delta} = \frac{\Delta}{(Km)(Km + \Delta)} \asymp \frac{\Delta}{(Km)^2}.$$

Next, consider the information distance. For exponential laws,

$$D_{\mathrm{KL}}\big(\,\mathrm{Exp}(\mu) \;\|\; \mathrm{Exp}(\mu')\big) = \log \frac{\mu}{\mu'} - 1 + \frac{\mu'}{\mu}.$$

Setting $\mu = Km + \Delta$, $\mu' = Km$, the divergence expands to $\frac{1}{2}(\Delta/(Km))^2 + O((\Delta/(Km))^3)$. Thus each play contributes $O((\Delta/(Km))^2)$ to the total KL. Over T rounds,

$$D_{\mathrm{KL}}(\mathsf{P} \;\|\; \mathsf{Q}) \;\leq\; c_0 T \Big(\tfrac{\Delta}{Km}\Big)^2$$

for some $c_0 > 0$.

By the Bretagnolle–Huber inequality,

$$\frac{\mathcal{R}_{\mathsf{P}}(T) + \mathcal{R}_{\mathsf{Q}}(T)}{2} \;\geq\; \Delta_{\mathrm{gap}} \cdot \frac{T}{2} \exp\Big(-D_{\mathrm{KL}}(\mathsf{P} \;\|\; \mathsf{Q})\Big).$$

Finally, choose $\Delta = \alpha Km/\sqrt{T}$ for a small constant $\alpha > 0$. Then $D_{\mathrm{KL}}(\mathsf{P} \;\|\; \mathsf{Q}) = O(1)$ and $\Delta_{\mathrm{gap}} \asymp (\alpha/(Km)) \cdot 1/\sqrt{T}$. Therefore

$$\frac{\mathcal{R}_{\mathsf{P}}(T) + \mathcal{R}_{\mathsf{Q}}(T)}{2} \;\geq\; c_1\sqrt{T},$$

for some $c_1 > 0$. Hence in at least one environment the regret satisfies $\mathcal{R}(T) \geq c\sqrt{T}$, completing the proof. □